

Rachad ANTONIUS

Sociologue, professeur, Département de sociologie, UQÀM

(2004)

Initiation au logiciel d'analyse statistique **SPSS.**

Laboratoires et exercices divers

Un document produit en version numérique par Jean-Marie Tremblay, bénévole,
professeur de sociologie au Cégep de Chicoutimi

Courriel: jean-marie_tremblay@uqac.ca

Site web pédagogique : <http://www.uqac.ca/jmt-sociologue/>

Dans le cadre de: "Les classiques des sciences sociales"

Une bibliothèque numérique fondée et dirigée par Jean-Marie Tremblay,
professeur de sociologie au Cégep de Chicoutimi

Site web: <http://classiques.uqac.ca/>

Une collection développée en collaboration avec la Bibliothèque
Paul-Émile-Boulet de l'Université du Québec à Chicoutimi

Site web: <http://bibliotheque.uqac.ca/>

VISIT...

LANZAROTE
Caliente.COM

Politique d'utilisation de la bibliothèque des Classiques

Toute reproduction et rediffusion de nos fichiers est interdite, même avec la mention de leur provenance, sans l'autorisation formelle, écrite, du fondateur des Classiques des sciences sociales, Jean-Marie Tremblay, sociologue.

Les fichiers des Classiques des sciences sociales ne peuvent sans autorisation formelle:

- être hébergés (en fichier ou page web, en totalité ou en partie) sur un serveur autre que celui des Classiques.
- servir de base de travail à un autre fichier modifié ensuite par tout autre moyen (couleur, police, mise en page, extraits, support, etc...),

Les fichiers (.html, .doc, .pdf., .rtf, .jpg, .gif) disponibles sur le site Les Classiques des sciences sociales sont la propriété des **Classiques des sciences sociales**, un organisme à but non lucratif composé exclusivement de bénévoles.

Ils sont disponibles pour une utilisation intellectuelle et personnelle et, en aucun cas, commerciale. Toute utilisation à des fins commerciales des fichiers sur ce site est strictement interdite et toute rediffusion est également strictement interdite.

L'accès à notre travail est libre et gratuit à tous les utilisateurs. C'est notre mission.

Jean-Marie Tremblay, sociologue
Fondateur et Président-directeur général,
[LES CLASSIQUES DES SCIENCES SOCIALES](#).

Cette édition électronique a été réalisée par Jean-Marie Tremblay, bénévole, professeur de sociologie au Cégep de Chicoutimi à partir de :

Rachad ANTONIUS

Initiation au logiciel d'analyse statistique SPSS. Laboratoires et exercices divers.

Montréal : Département de sociologie, UQÀM, 92 pp. Notes de cours pour le cours SOC 4206.

[Autorisation formelle accordée par l'auteur le 24 mai 2008 de diffuser cet article dans Les Classiques des sciences sociales.]



Courriel : antonius.rachad@uqam.ca

Polices de caractères utilisée :

Pour le texte: Times New Roman, 14 points.

Pour les citations : Times New Roman, 12 points.

Pour les notes de bas de page : Times New Roman, 12 points.

Édition électronique réalisée avec le traitement de textes Microsoft Word 2004 pour Macintosh.

Mise en page sur papier format : LETTRE (US letter), 8.5'' x 11''

Édition numérique réalisée le 24 mai 2008 à Chicoutimi, Ville de Saguenay, province de Québec, Canada.



UQÀM

Labos SPSS

Et exercices divers

Notes de cours
pour le cours

SOC 4206
Hiver 2004

Rachad Antonius
Département de sociologie

TABLE DES MATIÈRES

I – Labos SPSS

Labo 1	Introduction à SPSS	1.1
Labo 2	La syntaxe dans SPSS	2.1
Labo 3	Les mesures descriptives I – Catégories	3.1
Labo 4	Les mesures descriptives II – Variables quantitatives	4.1
Labo 5	La manipulation des données et des variables	5.1
Labo 6	Création de nouvelles variables calculées	6.1
Labo 7	Création d'un fichier de données	7.1
Labo 8	Les tableaux croisés	8.1
Labo 9	Les comparaisons de moyennes	9.1
Labo 10	La corrélation et la régression	10.1
Labo 11	Estimation et intervalles de confiance	11.1
Labo 12	Les tests T de validation d'une hypothèse	12.1
Labo 13	Les tests de Chi-deux	13.1

II - Divers

1.	Utilisation de Excel pour les calculs simples	14.1
2.	Calcul de la corrélation et la régression avec Excel	14.4
3.	Estimation – Notes	15.1
4.	Estimation- Exercices	15.5
5.	Exercices sur la courbe normale	16.1
6.	Format de l'examen final et modèle de réponses	17.1
7.	Questions sur le palmarès scolaire	18.1
8.	Diagrammes :	19.1
	Les deux branches des statistiques	
	L'inférence statistique	
	La validation d'hypothèse	
	Les procédures de mesure de l'association statistique	

LABO 1: INTRODUCTION À SPSS

1. Introduction

SPSS, dont le sigle anglais signifie Statistical Package for the Social Sciences, est un programme informatique d'analyse de données statistiques. Il permet de saisir des données, d'en faire des présentations résumées (tableaux, graphiques), de les organiser et surtout de les analyser. Il fonctionne sur les plateformes Macintosh et Windows, ainsi que sur les systèmes centraux tels UNIX. Nous ferons référence à trois types de documents SPSS :

- des fichiers de données (SPSS Data Editor),
- des fichiers de commandes permettant d'exécuter des procédures statistiques (SPSS Syntax Editor),
- et des fichiers de résultats où apparaissent les tableaux et les graphiques produits par SPSS (SPSS Viewer).

Dans ce qui suit nous apprendrons ce que sont ces fichiers et comment les utiliser.

Le programme SPSS est constamment mis à jour et amélioré, mais les différences entre les versions ne sont pas toujours majeures. Nous utiliserons ici la version 11, mais elle ne diffère pas beaucoup des versions 10 et 12 en ce qui concerne les besoins de ce cours. Les versions 8 et 9 diffèrent légèrement en apparence, mais les tableaux et les graphiques produits sont à peu près les mêmes que ceux obtenus avec les versions 10, 11 et 12. Il existe une version Française de SPSS. Mais les versions installées dans les labos de l'Université étant anglaises, nous ferons références à ces versions anglaises.

2. Démarrage de SPSS

Quand on démarre SPSS, on obtient soit un fichier de données vide, ou alors la **figure 1.1** que l'utilisateur a le choix de faire apparaître ou non au démarrage de SPSS.

Cette boîte de dialogue vous donne plusieurs choix, dont ouvrir un nouveau fichier, ouvrir un fichier existant, etc.. Si vous choisissez l'option **Open an existing data source** et puis **More Files...** qui est sélectionnée par défaut, et que vous cliquez **OK**, vous obtenez une liste de fichiers SPSS. En faisant défiler la fenêtre vers la droite, d'autres fichiers apparaissent. En regardant attentivement, vous verrez un fichier nommé **GSS93 subset**. Sélectionnez-le et ouvrez-le. Nous allons travailler beaucoup sur les données de ce fichier qui provient d'une enquête sociale générale (General Social Survey) entreprise aux États-Unis en 1993. Ce fichier est fourni avec les diverses versions de SPSS et il contient un assortiment de variables intéressantes à analyser. Il faut cependant noter que si les données sont bien réelles, l'échantillon de 1500 cas utilisé pour la construction de ce fichier n'est pas représentatif. Les conclusions qu'on en tirera ne reflètent donc pas les caractéristiques réelles de la population américaine en 1993.

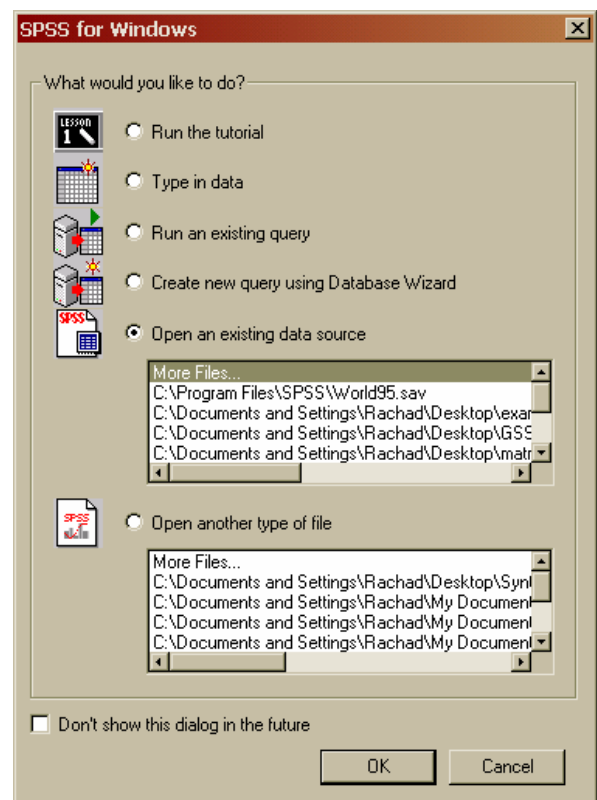


Fig. 1.1

En ouvrant le fichier de données **GSS93 subset**, on obtient la fenêtre suivante (Fig. 1.2).

	id	wrkstat	marital	agewed	sibs	childs	age	birthmo	zodiac	educ	degree
1	1	Working ful	divorced	20	3	1	43	May	Taurus	11	High schoo
2	2	Working ful	never marri	nap	2	0	44	August	Virgo	16	Bachelor
3	3	Working ful	divorced	25	2	0	43	February	Aquarius	16	Bachelor
4	4	Working pa	never marri	nap	4	0	45	NA	NA	15	High schoo
5	5	Retired	never marri	nap	1	0	78	October	Libra	17	Graduate
6	6	Retired	married	25	2	2	83	March	Pisces	11	High schoo
7	7	Working ful	married	22	2	2	55	October	Libra	12	High schoo
8	8	Retired	married	24	3	2	75	November	Sagittarius	12	High schoo
9	9	Working ful	divorced	22	1	2	31	July	Cancer	18	Graduate
10	10	Working pa	never marri	nap	1	0	54	March	Pisces	18	Graduate

Fig. 1.2.

Vous remarquerez que le nom du fichier apparaît au haut de cette fenêtre et que le terme **SPSS Data Editor** apparaît à sa suite. Ce terme désigne les fichiers de données. Dans ces fichiers, on peut saisir les données, les organiser et les transformer, d'où le terme *Data Editor*.

La fenêtre des fichiers de données comporte, au bas à gauche, deux onglets, étiquetés **Data View** et **Variable View**. Chacun de ces onglets correspond à l'un des deux affichages possibles de la fenêtre de données : soit les données elles-mêmes, ou alors la liste des variables ainsi que leurs caractéristiques. En cliquant sur un onglet ou sur l'autre, on passe d'un affichage à l'autre, en restant toujours dans le même document. La figure 1.2 montrée ci-haut correspond à l'affichage **Data View**. Ce sont les données elles-mêmes que l'on voit. Ces données sont organisées en lignes et en colonnes. Chaque ligne correspond à un cas, et chaque colonne à une variable. La première ligne comporte toutes les informations du questionnaire numéro 1, la deuxième ligne les informations du questionnaire numéro 2, et ainsi de suite. Mais attention : faites défiler cette feuille de données vers le bas jusqu'à la ligne 1500. Que voyez-vous dans la première case ? Le numéro qui apparaît n'est pas 1500. Vérifiez vous-même. On y lit : 1606. Ceci signifie que plusieurs questionnaires ont dû être abandonnés, sans doute parce qu'ils comportaient trop d'omissions, ou qu'ils étaient mal remplis. Pour avoir 1500 cas, on a dû se rendre au questionnaire 1606. Donc, 106 questionnaires ont été ignorés pendant la constitution de ce fichier.

Les colonnes correspondent à des variables. La première colonne identifie le cas, ou le questionnaire, qui reçoit un numéro. La 2^{ème} colonne correspond à la variable 'statut d'emploi'. Elle nous renseigne sur le statut de la personne, qui pourrait travailler à temps plein, ou à temps partiel, ou être retraitée, etc. La troisième colonne donne le statut matrimonial des individus de l'échantillon, et ainsi de suite.

Nous allons à présent expérimenter quelques manipulations de l'apparence du fichier. Vous êtes invités à les exécuter sur votre poste de travail.

1. L'affichage **Data View** peut faire apparaître soit les codes utilisés pour désigner les catégories des variables, soit les catégories elles-mêmes. Par exemple, on pourrait avoir dans la colonne de la variable sexe soit les codes 1 ou 2, ou les valeurs Hommes ou Femmes. Le changement d'une option à l'autre se fait en sélectionnant l'option **Value Labels** dans le menu **View** ou en la désélectionnant. Faites-en l'expérience et observez le résultat.
2. On peut aussi faire apparaître le nom complet d'une variable en positionnant le curseur au haut de la colonne correspondante, là où le nom bref apparaît. Ainsi, le nom complet de la variable **marital**

apparaît comme étant **Marital Status** quand on positionne le curseur dessus. Essayez, et trouvez aussi les noms complets des 3 ou 4 variables suivantes.

3. Élargissez une colonne en positionnant le curseur sur la ligne qui la sépare de la colonne suivante, puis en tirant vers la droite avec le bouton droit de la souris pressé.
4. **La commande Variables.** Sélectionnez la commande **Variables** dans le menu **Utilities**. Vous obtiendrez la fenêtre de la figure 1.3. On peut y lire les caractéristiques de chacune des variables en faisant défiler la liste des variables. En positionnant le curseur sur la variable **marital**, par exemple, vous verrez la signification de tous les codes utilisés :
 - 1 signifie Married
 - 2 signifie Widowed (i.e veuf ou veuve), etc.

Une explication est requise pour le mot **Type**. Il est suivi de codes de la forme : F4.1, ou encore F1. C'est le format dans lequel la variable est notée. F4.1 signifie que quatre espaces sont requis pour noter les valeurs de cette variable, dont un point et une décimale (le point occupe un espace). On pourra donc inscrire des valeurs telles que 28.3 qui prennent quatre espaces et qui comportent une décimale. Le format F2 signifie que la variable est notée par un nombre comportant deux chiffres, sans décimale. On peut aussi avoir des formats tels que A8 qui signifie que la variable est notée par 8 caractères qui n'ont pas de valeur numérique, mais le fichier **GSS93 subset** ne comporte pas de telles variables.

5. Vous remarquerez dans cette même fenêtre que le terme **Missing Values** apparaît avec chaque variable. Il désigne les valeurs manquantes, qui sont spécifiées. Ces valeurs sont utilisées pour coder des situations telles que 'Le répondant refuse de répondre' ou 'La question ne s'applique pas'. Dans de telles situations, on ne veut pas que les valeurs correspondantes soient prises en considération dans les calculs statistiques. La mention **Missing Values** nous indique que ces valeurs ne seront pas prises en considération dans les calculs. Nous verrons au labo 9 comment définir les variables et spécifier les valeurs manquantes.

Le terme **Measurement Level** désigne l'échelle de mesure utilisée pour cette variable (nominale, ordinale ou échelle quantitative, notions vues au premier cours).

6. **La commande File Info.** Sélectionnez la commande **File Info** sous le menu **Utilities**. Vous verrez une nouvelle fenêtre apparaître, qui comporte toutes les informations vues dans la fenêtre **Variables** mentionnée ci-haut. Cette fenêtre est intitulée **Output1** et elle est de type SPSS Viewer, qui est le type de fichier qui comporte les tableaux et graphiques produits par SPSS. L'avantage de produire ces informations par la commande **File Info**, c'est qu'on peut copier toutes ces informations d'un seul coup et les coller dans un document Word (ou tout autre traitement de texte) et les faire imprimer en tout ou en partie. Essayez cette procédure : cliquez une fois sur les informations produites par **File Info** : une bordure apparaît, indiquant que cette information est sélectionnée. Copiez et collez dans un document Word.

Chaque fois qu'on donne une commande SPSS, le résultat est affiché dans une fenêtre de type **SPSS Viewer**. Les résultats des commandes suivantes sont affichés dans le même fichier, à la suite des résultats déjà produits. On peut enregistrer ce fichier de résultats en lui attribuant un nom de notre choix. On peut aussi sélectionner n'importe quel résultat apparaissant dans ce fichier, puis le copier et le coller dans un document Word.

Vous aurez sans doute remarqué que la fenêtre de résultats est divisée verticalement en deux. Le côté gauche de la fenêtre comprend une sorte de plan, ou de table des matières des résultats produits (le terme utilisé par SPSS pour désigner cette partie est : *Document map*). Quand le fichier de résultats comprend de nombreux éléments, le *Document map* permet de repérer rapidement un résultat et de le visionner.

Nous n'avons pas encore parlé des fichiers de type SPSS Syntax Editor. Ceci fera l'objet du labo 2.

Exercice 1.1

Ouvrez le fichier intitulé **Road Construction Bids** qui est fourni avec SPSS. Pour cela, cliquez sur **Open → Data**, repérez ce fichier en faisant défiler la fenêtre vers la droite, et cliquez deux fois dessus.

Produisez les informations sur les variables de ce fichier (**File Info**), puis copiez-les dans un document Word.

IMPORTANT : Incluez dans ce document un en-tête qui comprend votre nom, le numéro du labo, la date, ainsi que la pagination. Cette opération a été montrée en classe. Tous les documents que vous allez produire dans ce cours doivent comporter un tel en-tête.

Enregistrez ce document sur votre disquette afin de le faire imprimer si nécessaire.

Labo 2 La syntaxe dans SPSS

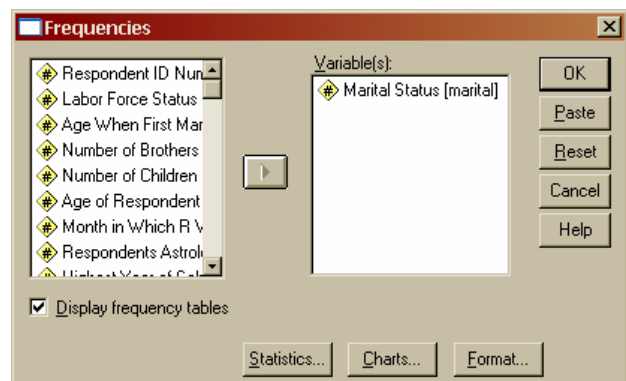
Les commandes dans SPSS peuvent être données de deux façons : soit en cliquant dans le menu approprié, soit en écrivant la commande dans un langage spécial qui doit obéir à une syntaxe très précise, et qui apparaît dans une fenêtre de type **SPSS Syntax Editor**. Certaines commandes ne peuvent être données que dans le langage de la syntaxe, mais nous n'aurons pas à traiter de telles commandes dans ce cours.

Rappelez-vous qu'il y a trois sortes de fenêtres dans SPSS :

1. Celles où les données apparaissent, appelées **SPSS Data Editor**, qui ont elles-mêmes deux modes d'affichage : **Data View** permet de voir les données elles-mêmes, et **Variable View** permet de voir les propriétés de chacune des variables, qui sont listées sur le même écran, chaque variable occupant une ligne. On passe de l'un de ces deux affichages à l'autre en cliquant sur l'onglet approprié au bas de l'écran, à gauche. On ne peut ouvrir qu'une seule fenêtre de données à la fois.
2. Celles où les tableaux et les graphiques apparaissent, appelées **SPSS Viewer**. On peut avoir plusieurs fenêtres de type Viewer ouvertes à la fois.
3. Et celles où la syntaxe apparaît, appelées **SPSS syntax Editor**. On peut avoir plusieurs fenêtres de type Syntax Editor ouvertes à la fois.

On peut sauvegarder chacune de ses fenêtres et lui donner un nom. Ainsi, si vous avez produit des tableaux de fréquence, vous pouvez cliquer la commande **Save** et sauvegarder votre document sous le titre, disons, de 'Labo 4_Votre_nom_de_famille'.

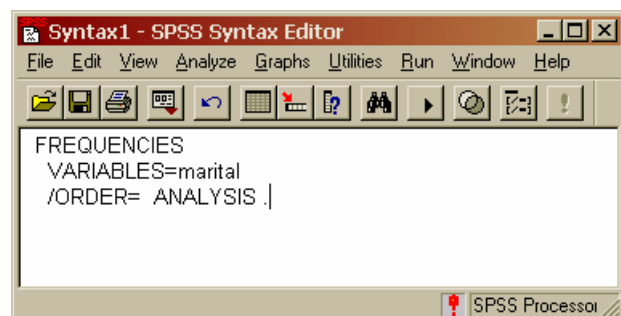
Il y a deux façons d'écrire une syntaxe. Soit que vous la dactylographiez (les utilisateurs réguliers de SPSS préfèrent cette méthode) ou encore que vous demandiez à SPSS de l'écrire pour vous. En effet, lorsque vous donnez une commande par menus, vous avez toujours l'option de cliquer **Paste** plutôt que **OK**, ce qui a pour effet de coller la syntaxe correspondante dans la fenêtre de la syntaxe. Regardez la figure ci-contre. C'est la boîte de dialogue ('Dialogue Box', en anglais) de la commande **Frequencies**. On a placé la variable **Marital Status**, qui provient du fichier **GSS93 subset**, dans l'espace des variables à traiter dans cette commande. Si on clique **OK**, on obtiendra le tableau de fréquences de l'état matrimonial des répondants. Mais si on clique **Paste**, une nouvelle fenêtre s'ouvre, illustrée ci-bas.



NOTE :

Il est suggéré d'exécuter les commandes illustrées en même temps que vous les lisez, afin de bien les comprendre.

On voit dans cette fenêtre la structure de la syntaxe : la commande utilisée est d'abord indiquée (FREQUENCIES). Sur la ligne suivante, en retrait, il est indiqué la liste des variables à laquelle cette commande s'applique. La troisième ligne indique l'ordre dans lequel les tableaux vont paraître, au cas où il y aurait plusieurs variables à traiter.



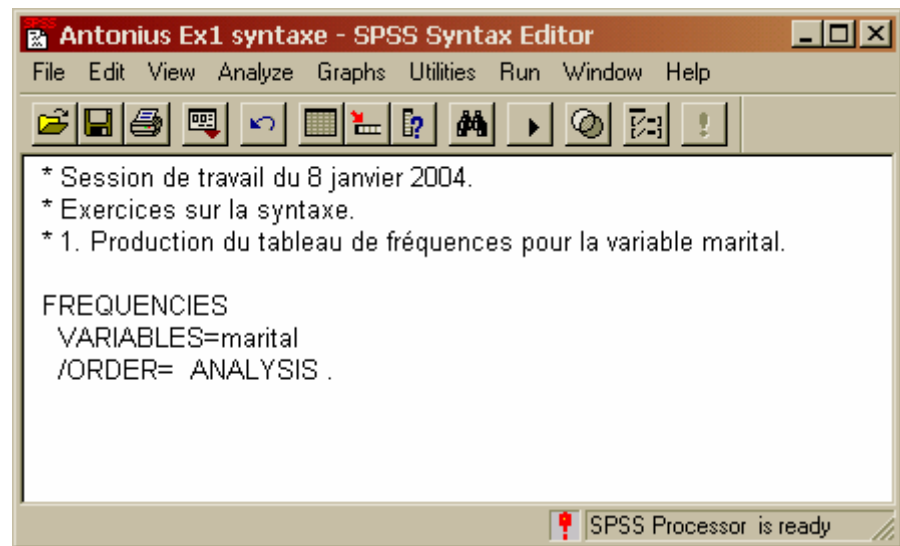
Vous n'aurez pas à écrire la syntaxe des commandes vous-mêmes dans ce cours, mais il faudrait apprendre à la faire écrire par SPSS, comme nous venons de le faire.

En effet, la syntaxe a plusieurs avantages :

1. On peut la sauvegarder et refaire les analyses statistiques plus tard, ou les refaire en utilisant d'autres données qui comportent les mêmes variables, ou encore une partie des données.
2. Elle permet de tenir une sorte de 'journal' de toutes les commandes qui ont été utilisées dans une session de travail.
3. On peut la copier, la coller, la modifier, comme on fait pour n'importe quel texte. On peut ajouter des variables à la liste des variables traitées.
4. On peut écrire des commentaires avant ou après la syntaxe, pour expliquer ce qu'on voulait faire, ou pourquoi on l'a fait, ou pour tout autre commentaire. Cependant, chaque ligne de commentaires doit être précédée d'un astérix * (qu'on obtient en tapant majuscule 8) qui indique à SPSS que ceci est un commentaire et qu'il ne faut pas le traiter comme une commande. Il vaut mieux mettre un point à la fin d'un commentaire ou encore laisser une ligne blanche.
5. Les commandes écrites en syntaxe ne sont exécutées que lorsque vous demandez à SPSS de le faire, soit en cliquant le menu **Run → All** dans la fenêtre de la syntaxe, ou encore en sélectionnant une commande puis en cliquant sur le petit triangle noir qui se trouve parmi les icônes au haut de la fenêtre de la syntaxe.

Exemple de commentaire :

Les trois premières lignes de cette fenêtre commencent par un astérix et sont donc considérées comme des commentaires et non pas des commandes. Vous remarquerez aussi que ce document de syntaxe a été sauvegardé sous le nom de **Antonius Ex1 syntaxe**. Il est fort utile d'ajouter votre nom aux documents que vous produisez quand vous devez les remettre sous forme électronique, afin que le correcteur sache qui les a produits.



EXERCICE 2.1

Produisez la syntaxe nécessaire pour obtenir les tableaux de fréquence des variables **Labor Force Status** (wrkstat) et **Number of Children** (childs). Écrivez un commentaire explicatif et sauvegardez le document de syntaxe que vous avez produit. Prenez l'habitude d'ajouter une ligne vide après chaque commande collée par SPSS, afin que vous puissiez différencier une commande de la commande suivante.....

LABO 3: LES PROCÉDURES DESCRIPTIVES I - CATÉGORIES

Le but de ce labo est de vous familiariser avec les procédures les plus communes pour décrire des données avec SPSS. Nous travaillons avec le fichier de données **GSS93 subset** qui est fourni avec le programme SPSS. Afin de se familiariser avec les procédures disponibles dans SPSS, nous répondrons à des questions telles que :

- Quel est l'âge moyen des personnes à leur premier mariage ?
- Quel est l'âge moyen des hommes dans cet échantillon ?
- Quel est l'âge moyen des femmes ?
- Peut-on donner une représentation visuelle de la distribution de la variable « âge » ?
- Quelle est la proportion de personnes qui favorisent la peine de mort ?

Il y a quatre commandes qui produisent des mesures descriptives. Elles se trouvent toutes dans le menu **Analyze** → **Descriptive statistics** tel qu'illustré par la figure 3.1. Ces procédures sont :

La commande **Frequencies...** qui produit des tableaux de fréquences mais d'autres mesures aussi,

La commande **Descriptives...** qui produit des mesures quantitatives,

La commande **Explore...** qui produit diverses mesures d'un seul coup, pour l'ensemble des données ou encore pour des sous-groupes qui peuvent alors être comparés,

et la commande **Crosstabs** qui produit des tableaux croisés de deux variables ou plus.

(Nous n'utiliserons pas la commande **Ratio...** pour le moment).

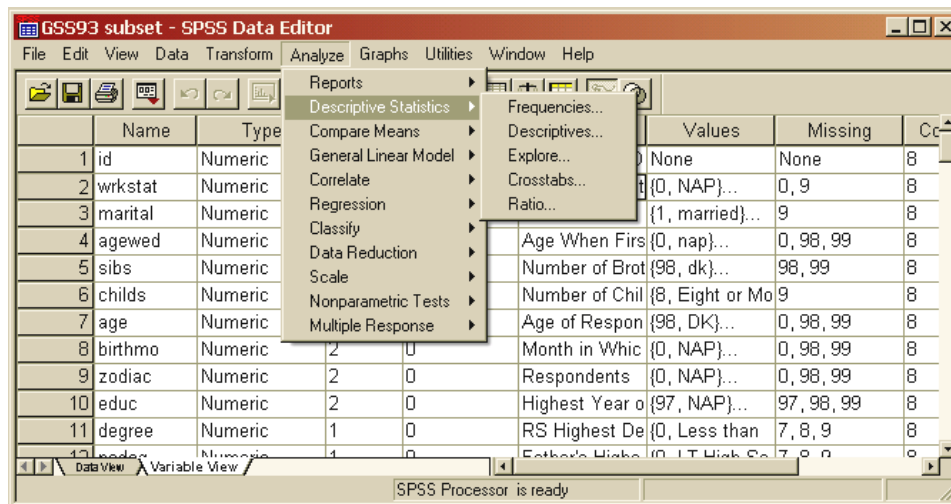


Figure 3.1

Pour chacune de ces commandes, il faut spécifier les variables à analyser, ainsi que certaines options offertes par SPSS. Ce laboratoire est de nature exploratoire : il vous permettra de vous familiariser avec ces diverses procédures. Rappelez-vous cependant que l'échelle de mesure utilisée pour une variable détermine les procédures que l'on peut lui appliquer : il ne sert à rien de calculer une moyenne quand la variable est qualitative, par exemple. Les tableaux de fréquences de la procédure **Frequencies...** sont appropriés quand on a un nombre restreint de catégories, et qu'on veut mesurer leur importance relative ou absolue. Par contre, cette même procédure offre de nombreuses options intéressantes pour les variables

quantitatives. Les procédures **Descriptives...** et **Explore...** ne sont applicables que pour les variables quantitatives.

Attention: SPSS est un programme d'analyse statistique puissant, qui offre une grande étendue de possibilités. Nous n'en utiliserons qu'une petite partie. Il vous faudra donc spécifier uniquement les options que vous connaissez, et ne pas modifier celles que vous ne connaissez pas et qui sont offertes par défaut par SPSS. Si vous obtenez accidentellement des tableaux que vous ne savez pas interpréter, ne les utilisez pas dans les résumés d'analyses que vous ferez.

Variables qualitatives, ou quantitatives comportant un petit nombre de catégories

La commande *Frequencies...*

1. Sélectionnez la commande **Frequencies...** montrée ci-haut.

Vous obtenez la boîte de dialogue illustrée à la figure 3.1. Cette procédure est utile quand les variables sont qualitatives, mais elles sont aussi très utiles quand la variable est quantitative mais qu'elle a été regroupée en un nombre restreint de catégories, comme par exemple pour la variable **Age Categories** [**agecat4**] qui se trouve vers la fin de la liste de variables

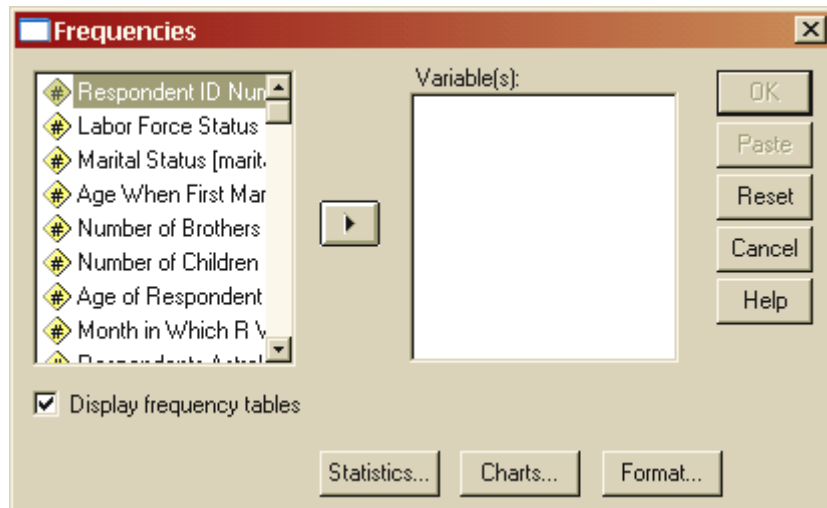


Figure 3.2

Toutes les variables du fichier sont inscrites dans la partie gauche de cette boîte de dialogue. Pour obtenir le tableau de fréquence d'une variable, il faut la sélectionner, puis la placer dans l'espace prévu à droite en cliquant sur le bouton contenant une mini-flèche. Remarque qu'il y a plusieurs boutons permettant de spécifier des options.

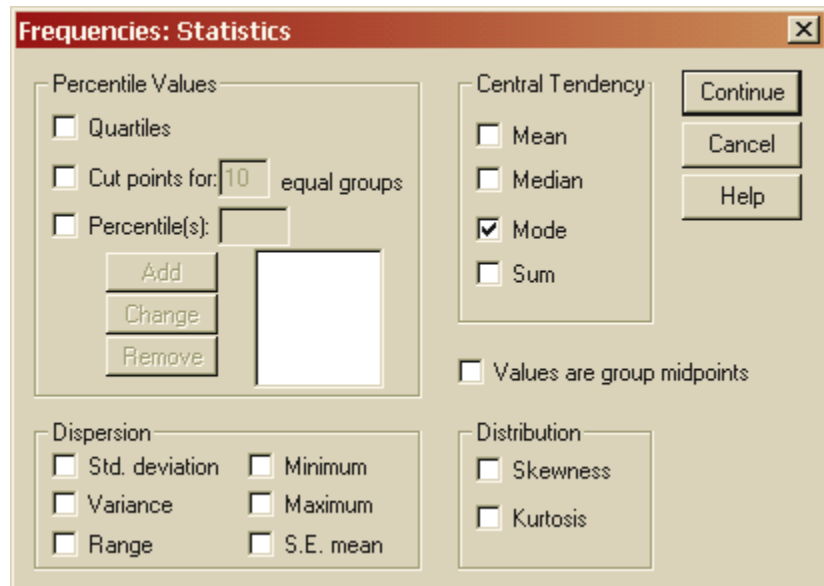
2. Sélectionnez les variables **Marital Status** et **Age Categories** (attention : pas **Age of Respondent** qui n'est pas regroupée et qui comporte un trop grand nombre de catégories) et placez-les dans l'espace prévu à droite.
Laissez le petit carré de l'option **Display frequency table** sélectionné.

3. Cliquez maintenant sur le bouton **Statistics...**

Vous obtenez la boîte de dialogue illustrée à la Figure 3.3.

Il y a quatre sections dans cette boîte de dialogue, chacune permettant un type de mesure descriptives : des mesures de position telles que les quartiles ou les percentiles, des mesures de tendance centrale, des mesures de dispersion, et des mesures qui décrivent la distribution dans son ensemble. Revoyez les définitions de ces termes vues au début du cours. Si la variable est qualitative, seul le **Mode** sera utile parmi ces mesures.

Figure 3.3



4. Cliquez **Continue**, vous reviendrez à la boîte de dialogue précédente.
5. Cliquez sur le bouton **Charts...** . Vous obtenez la figure 3.4.

On a le choix entre plusieurs type de graphiques. Choisissez **Bar charts** et cliquez **Continue**.

6. Dans la boîte de dialogue initiale de la commande **Frequencies**, cliquez sur **Paste**. Cette opération inscrit la commande dans la fenêtre du **Syntax Editor**. Vous devriez obtenir la commande suivante :

```
FREQUENCIES
  VARIABLES=marital agecat4
  /STATISTICS=MODE
  /BARCHART FREQ
  /ORDER= ANALYSIS .
```

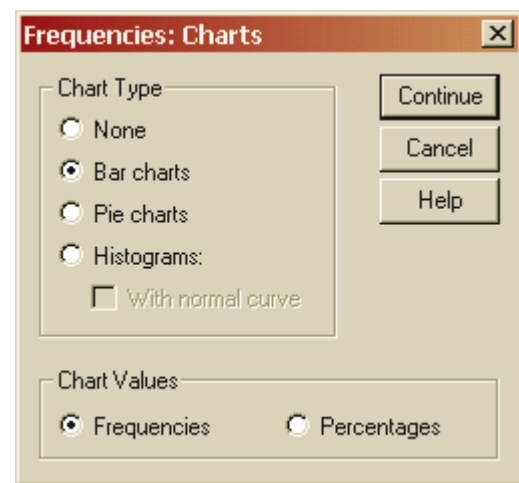


Figure 3.4

On voit ici les composantes de cette commande : La commande principale (FREQUENCIES) est suivies des sous-commandes qui spécifient les options possibles : la sous-commande VARIABLES (obligatoire) qui permet de spécifier les variables que l'on veut décrire, la sous-commande STATISTICS qui spécifie qu'on souhaite que le mode soit donné, une sous-commande pour les graphiques (BARCHART) et enfin une sous-commande qui spécifie l'ordre dans lequel les résultats vont apparaître.

Évidemment, on peut toujours cliquer **OK** plutôt que **Paste**. Dans ce cas, la commande est exécutée directement, sans que la syntaxe ne soit donnée.

Exercice 3.1

- a) Exécutez la commande donnée ci-haut et écrivez une phrase complète pour chacune des variables, qui décrit sa distribution en donnant les pourcentages appropriés.
- b) Refaites le même exercice en sélectionnant plutôt **Pie Charts**, puis une autre fois avec **Histograms**. Écrivez quelques lignes pour dire les avantages ou inconvénients comparatifs de ces trois types de graphiques pour représenter des variables qualitatives.

Lecture et interprétation des résultats SPSS

Les procédures expliquées ci-haut donnent un tableau et un graphique pour chaque variable, mais ces résultats sont généralement accompagnés d'un tableau qui liste toutes les variables, ainsi que le nombre total de cas valides que l'on a pour chacune. Ce premier tableau ne nous renseigne pas sur la distribution des variables proprement dites, mais il est important de l'examiner afin de connaître l'importance relative du nombre de données manquantes.

Les tableaux de fréquences. Ils comportent cinq colonnes. Produisez les tableaux de fréquences pour la variable **marital** et examinez-les et décrivez ce que chaque colonne contient. Quelle différence voyez-vous entre les colonnes **Percent** et **Valid percent** ? Quant à la colonne **Cumulative percentage**, elle n'est utile que pour les variables mesurées par une échelle ordinale ou d'intervalle ou de ratio. Elle donne le pourcentage cumulatifs des diverses catégories.

Les lignes d'un tableau de fréquences. Les premières lignes correspondent aux diverses catégories du tableau, et elles sont suivies d'une ligne qui donne le nombre total de réponses valides. Observez bien le pourcentage total de réponses valides. Si beaucoup de données manquent, il faut se demander pourquoi, car une grande proportion de données manquantes pourrait rendre toute généralisation problématique. Les lignes suivantes font justement le décompte des données manquantes, en les ventilant selon la raison pour laquelle elles sont manquantes quand cela est possible. En général, on utilise trois catégories de données manquantes. Les sigles anglais suivants sont utilisés dans les fichiers d'exemples de SPSS :

DK (Don't Know) ; quand la personne interrogée dit ne pas connaître la réponse.

NA (No Answer) ; quand la personne n'a pas répondu du tout.

NAP (Not Applicable) ; quand la question ne s'applique pas. Par exemple, si la question est : quel est l'âge de votre enfant aîné, et que la personne n'a pas d'enfants.

Les diagrammes en bâtons. Examinez la représentation graphique en bâtons de la variable **agecat4**. Il s'agit de l'âge regroupé en quatre catégories. Vous aurez remarqué sans doute que la barre la plus haute désigne la catégorie « 50 ans et plus ». Ceci est dû au fait que cette catégorie recouvre une étendue de près de 40 ans, alors que d'autres catégories ne couvrent qu'une étendue de 10 ans. Ceci donne une image un peu déformée de la distribution. Des catégories d'étendue égale permettent de faire de moyennes, et donnent une meilleure représentation de la distribution. Cependant, il y a parfois de bonnes raisons de regrouper les âges en catégories inégales. C'est la problématique à laquelle on s'adresse qui peut nous amener à préférer un regroupement en catégories inégales ou pas.

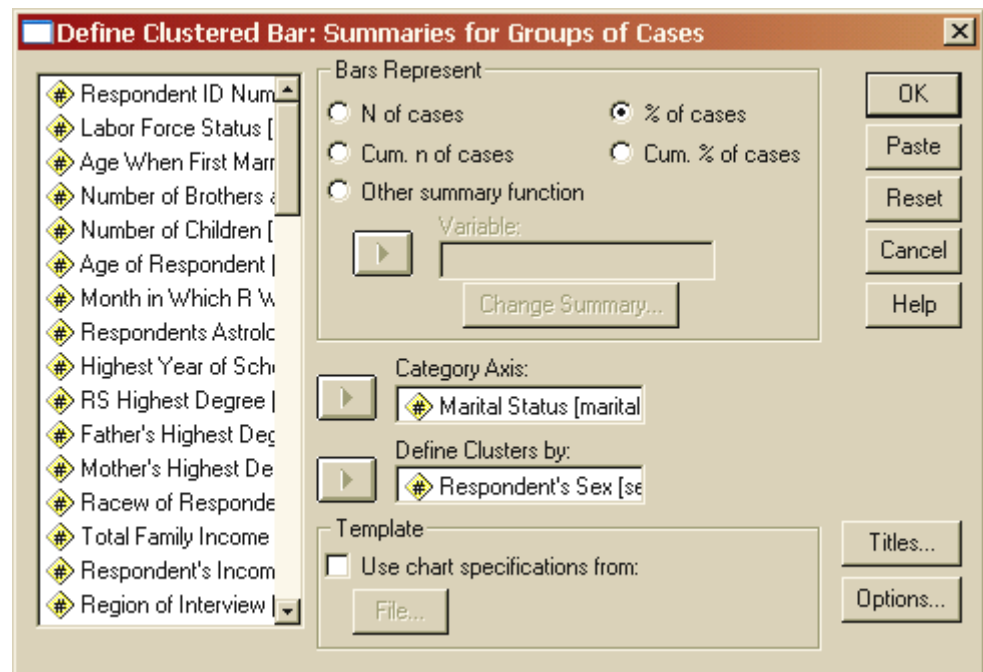
Productions de tableaux et graphiques ventilés par sous-groupe

Quand on produit des diagrammes en bâtons, il est souvent souhaitable de ventiler les résultats en sous-groupes. Ceci signifie que plutôt qu'obtenir les pourcentages des diverses catégories d'une variable pour l'ensemble de vos données, vous les obtenez d'abord pour les hommes versus les femmes, ou encore pour ceux qui ont un diplôme universitaire versus ceux qui n'en ont pas, et ainsi de suite. Cette procédure s'appelle **Clustered Bar Charts** dans la version anglaise de SPSS. On verra aussi comment obtenir des tableaux de fréquences ventilés de la même façon.

Pour produire un diagramme en bâtons ventilé (**Clustered Bar Chart**), suivre les étapes suivantes.

1. Sélectionnez **Bar...** sous le menu **Graphs**.
2. Dans la boîte de dialogue qui en résulte, sélectionnez **Clustered** et **Summaries for groups of cases**.
3. Cliquez **Define**. Vous obtenez la boîte de dialogue illustrée à la figure 3.5.
4. Placez la variable **Marital status** dans l'espace intitulé **Category Axis**, et placez la variable **Respondent's sex** dans l'espace désigné par **Define Clusters by**, tel qu'illustré dans la figure 3.5.
5. Au haut de cette boîte de dialogue, assurez-vous que pour l'option **Bar represents**, vous avez sélectionné **% of cases** plutôt que **N of cases**. La raison de ce choix est la suivante : comme il y a beaucoup plus de femmes que d'hommes dans notre fichier de données, des diagrammes en bâtons qui représenteraient le nombre de cas dans chaque catégorie donneraient une fausse impression de l'importance *relative* des catégories. Tandis que les pourcentages permettraient de comparer le pourcentage d'hommes dans une catégorie avec le pourcentage de femmes dans la même catégorie.
6. Cliquez le bouton **Options** et dé-sélectionnez le choix de faire apparaître les catégories relatives aux données manquantes (le libellé anglais est : **Display groups defined by missing values**).
7. Cliquez **OK** pour exécuter la commande directement, ou **Paste** pour obtenir la syntaxe correspondante.

Figure 3.5



La syntaxe obtenue est :

```
GRAPH
  /BAR(GROUPED)=PCT BY marital BY sex.
```

Vous obtenez un diagramme en bâtons où il devient évident que les hommes se retrouvent dans la catégorie “mariés” en plus grand pourcentage que les femmes, mais que ces dernières sont relativement plus nombreuses dans la catégorie « veufs/veuves ». Outre que l'échantillon que nous avons n'est sans doute pas représentatif, ce phénomène est dû au fait que les femmes ont tendance à vivre plus longtemps que les hommes. Il y a donc plus de chances qu'elles se retrouvent veuves.

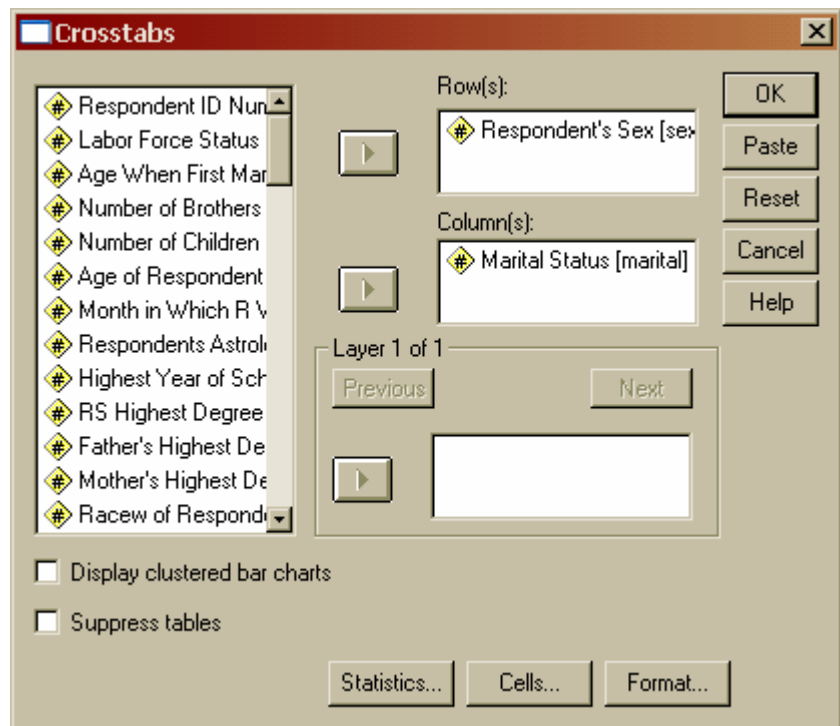
Exercice 3.2

Modifiez la syntaxe précédente pour obtenir le diagramme en bâtons pour la variable **marital**, mais ventilé en fonction de la possession ou nom d'un diplôme universitaire, et écrivez quelques lignes pour interpréter le diagramme obtenu. **Attention** : il y a plusieurs variables qui traitent du niveau d'éducation. Choisissez la bonne : elle n'a que deux catégories.

Tableaux de fréquences ventilés

C'est en utilisant la commande **Analyze→ Descriptive Statistics→ Crosstabs** qu'on obtient des tableaux de fréquences ventilés en sous-catégories. Cette procédure statistique est utilisée à sa pleine capacité dans le cadre de l'analyse de la relation entre deux variables, abordée plus loin dans le cours. Mais nous pouvons ici nous y référer pour produire les tableaux de fréquences ventilés par sous-groupes, en nous en tenant à une lecture directe des tableaux. Ceci est fait de la façon suivante.

1. Sélectionnez la commande **Analyze→ Descriptive Statistics→ Crosstabs**. Vous obtenez la boîte de dialogue illustrée à la **figure 3.6** ci-contre.
2. Dans la liste des variables, sélectionnez la variable **Respondant's sex** et placez-la dans l'espace réservé pour les lignes (Rows), et sélectionnez la variable **Marital Status** et placez-la dans l'espace réservé pour les colonnes, tel qu'illustré dans la figure 3.6.
3. Vous pouvez cocher la case **Display cluster bar charts**, mais vous ne pourrez pas modifier les options par défaut du diagramme : vous obtiendrez les colonnes du graphique qui représentent les fréquences (et non les pourcentages) et les catégories des données manquantes apparaîtront dans le graphique. Si vous avez vraiment besoin des graphiques, il vaut mieux les produire par la commande **Graphs** comme montré plus haut.
4. Cliquez le bouton **Cells** et assurez-vous que les cases pour **Rows** et pour **Observed** sont cochées. Ne cochez aucune des cases du bouton **Statistics**. Nous verrons ces options plus tard dans le cours, quand nous parlerons d'inférence statistique.



Cliquez **OK** ou encore **Paste** si vous préférez travailler avec la syntaxe. Vous devriez obtenir la syntaxe suivante :

```
CROSSTABS  
  /TABLES=sex BY marital  
  /FORMAT= AVALUE TABLES  
  /CELLS= COUNT ROW .
```

Si vous l'exécutez, vous obtiendrez le tableau de fréquence ventilé.

Exercice 3. 4

Exécuter les commandes expliquées précédemment, et examinez le tableau qui en résulte. Répondez aux questions suivantes.

- a) Quel est le pourcentage d'hommes mariés ?
- b) Quel est le pourcentage de femmes mariées ?
- c) Quel est le pourcentage de personnes mariées ?
- d) Quel est le pourcentage d'hommes veufs ?
- e) Quel est le pourcentage de femmes veuves ?
- f) Quel est le pourcentage de personnes veuves ?

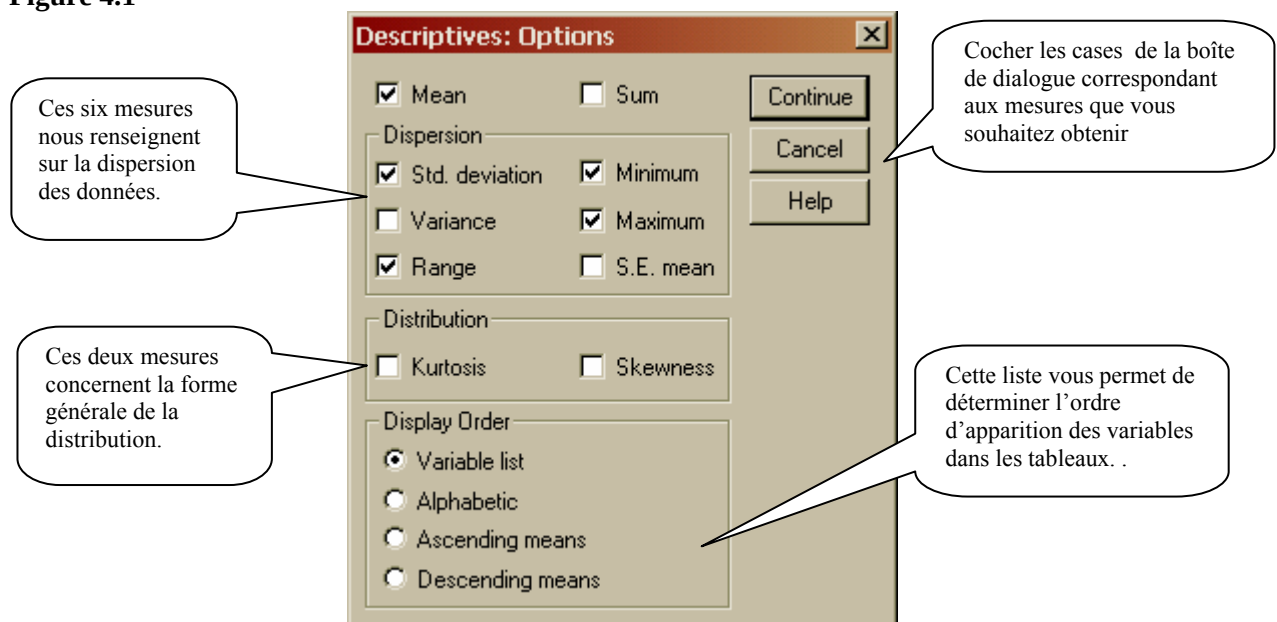
LABO 4: LES PROCÉDURES DESCRIPTIVES II - VARIABLES QUANTITATIVES

La commande Descriptives...

Cette commande n'est appropriée que pour les variables quantitatives, préférablement mesurées avec une échelle d'intervalle ou de ratio (**scale** dans la terminologie de SPSS). Nous allons l'illustrer avec un exemple que vous êtes invité à exécuter sur votre poste de travail.

1. Sélectionnez la commande **Descriptives...** (Analyze → Descriptive Statistics → Descriptives...).
2. Placez les variables que vous voulez analyser dans l'espace désigné par **Variables** du côté droit de la boîte de dialogue obtenue. Nous allons le faire pour les variables **Age of Respondent**, et **Age when First Married**.
3. Cliquez sur le bouton **Options** pour spécifier les statistiques que vous souhaitez obtenir. Vous obtenez la boîte de dialogue illustrée dans la figure 4.1. Remarquez que vous ne pouvez pas obtenir de graphiques par l'entremise de cette commande.

Figure 4.1



4. Cliquez **Continue**. Vous revenez à la boîte de dialogue principale de la commande **Descriptives...**
5. Cliquez **OK**, ou encore **Paste** si vous voulez travailler avec la syntaxe. Dans ce dernier cas, la syntaxe obtenue est :

```
DESCRIPTIVES
  VARIABLES=agewed age
  /STATISTICS=MEAN STDDEV RANGE MIN MAX .
```

Le résultat de l'exécution de la commande **Descriptives ...** est un tableau qui comporte toutes les mesures sélectionnées dans les **Options**. Dans l'exemple présent, vous obtenez :

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation
Age When First Married	1202	45	13	58	22,79	5,033
Age of Respondent	1495	71	18	89	46,23	17,418
Valid N (listwise)	1199					

Examinez la première colonne de ce tableau. Les variables analysées sont listées, et le nombre de réponses valides pour chaque variable est donné à la colonne 2. Mais la dernière ligne de la colonne 1 comporte le terme : **Valid N (listwise)**. Le nombre 1199 donné à la dernière ligne de la colonne 2 est le nombre de cas pour lesquels on a des données valide pour chacune des variables listées. On a donc 1199 cas pour lesquels on a à la fois l'âge du répondant et son âge au premier mariage.

La commande Explore...

La commande **Explore...** (Analyze → Descriptive Statistics → Explore...) s'applique elle aussi aux variables quantitatives uniquement. Elle nous permet d'obtenir une variété de mesures descriptives, ainsi que quelques mesures utilisées dans l'inférence statistique. Comme son nom le laisse supposer, elle est très utile dans une démarche exploratoire visant à se faire une idée générale de la distribution d'une variable. Elle permet en outre de traiter plusieurs variables d'un seul coup, et aussi de ventiler les données en fonction de sous-groupes définis par une variable qualitative (par exemple d'obtenir les mesures souhaitées séparément pour les hommes et les femmes). Nous allons illustrer ces usages par un exemple.

1. Cliquez sur la commande **Explore...** (Analyze → Descriptive Statistics → Explore...). Vous obtenez la boîte de dialogue illustrée à la figure 4.2.

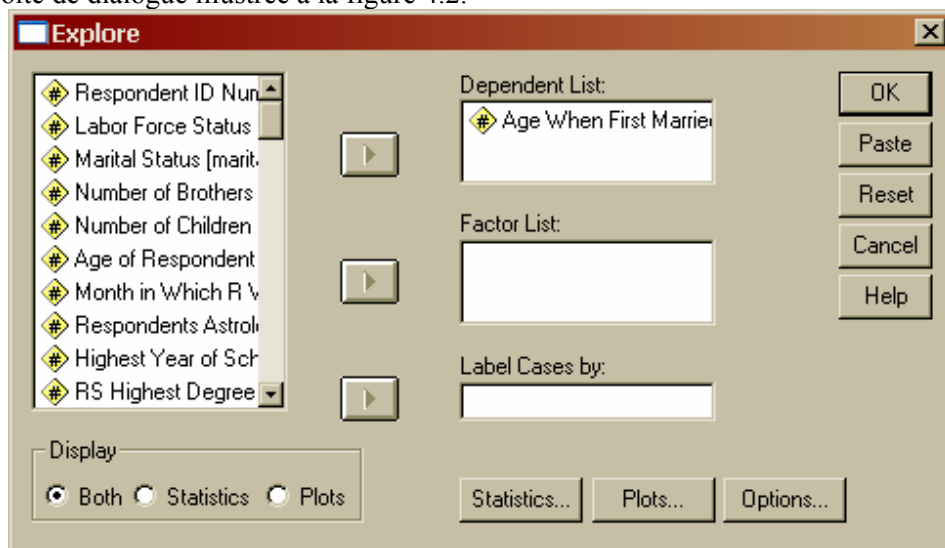


Figure 4.2

2. Sélectionnez la variable **Age When First Married** et placez-la dans l'espace désigné par le terme **Dependent List**, tel qu'illustré ci-haut. Laissez les autres espaces vides pour le moment.
3. Cliquez **OK** ou **Paste**. Nous examinerons la syntaxe un peu plus loin. Observez pour le moment les deux tableaux obtenus.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Age When First Married	1202	80.1%	298	19.9%	1500	100.0%

Descriptives

			Statistic	Std. Error
Age When First Married	Mean		22.79	.145
	95% Confidence Interval for Mean	Lower Bound	22.51	
		Upper Bound	23.08	
	5% Trimmed Mean		22.40	
	Median		22.00	
	Variance		25.331	
	Std. Deviation		5.033	
	Minimum		13	
	Maximum		58	
	Range		45	
	Interquartile Range		6.00	
	Skewness		1.658	.071
	Kurtosis		5.382	.141

Lecture des tableaux obtenus par la commande Explore...

Le premier tableau concerne le nombre de données valides, et il est plus utile quand nous traitons plusieurs variables simultanément.

Les mesures du deuxième tableau ont été expliquées dans la partie théorique du cours, sauf les cellules qui ont été ombragées et que nous ignorerons pour le moment : elles concernent l'inférence statistique qui sera étudiée plus loin dans le cours. Rappelez vous que la mesure **5% Trimmed Mean** est la moyenne de la variable après avoir éliminé les 5 % des cas les plus extrêmes. Cette mesure est utile quand des cas exceptionnels haussent ou baissent la moyenne de façon marquée. Mais il n'est pas toujours sage d'éliminer ces cas extrêmes : en fonction des questions que l'on se pose, ces cas pourraient être très révélateurs, mais ils pourraient aussi affecter démesurément les tendances centrales des données. C'est pourquoi il est toujours conseillé d'examiner tant la moyenne générale, que la moyenne quand les données extrêmes ont été supprimées et se poser la question de la signification de la différence entre les deux. Dans ce cas-ci, la différence est petite, et elle est due au fait de quelques premiers mariages tardifs : par exemple une personne s'est mariée pour la première fois à 58 ans, ce qui est l'âge maximum du premier mariage pour les données de ce fichier. La mesure **Skewness** est une mesure d'asymétrie, et elle est égale à 0 quand la mesure est parfaitement symétrique. Si la courbe représentant la distribution est étirée vers la droite, l'asymétrie est positive. Si elle est étirée vers la gauche, l'asymétrie est négative. La mesure **Kurtosis** est une mesure de la relative 'platitude' de la courbe représentant la distribution. Une courbe normale a une **Kurtosis** égale à 0, et cette mesure est positive quand la courbe est plus pointue qu'une courbe normale, et négative quand elle est plus plate qu'une courbe normale. Ces deux mesures ne sont utiles que pour comparer des distributions. Pour les fins de ce cours, l'inspection visuelle de la courbe sera probablement plus 'parlante' que ces mesures.

Les options de la commande Explore...

Vous aurez peut-être remarqué que dans la boîte de dialogue de la commande Explore..., on peut cocher une option qui permet de faire paraître soit des mesures statistiques uniquement, ou des graphiques, ou les deux. Le choix par défaut est **Both** (les deux).

De plus, des boutons spécifiques nous permettent de spécifier d'autres options. Examinons-les.

Les options du bouton Statistics

Quand on clique sur le bouton **Statistics**, on obtient la boîte de dialogue de la figure 4.3 illustrée ci-contre.

Si on garde l'option **Descriptives** cochée (elle l'est par défaut), on obtient toutes les statistiques obtenues précédemment. Le 95 % que l'on voit dans une case réfère à l'inférence statistique et sera discuté dans une étape ultérieure du cours.

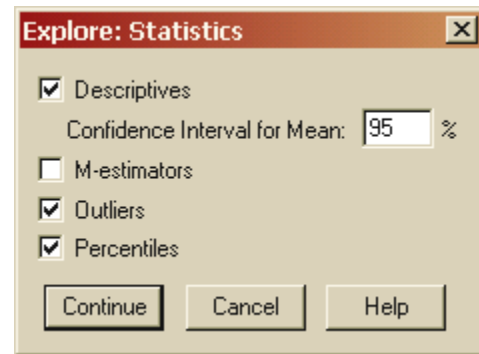


Figure 4.3

L'option **Outliers** établit une liste des 10 valeurs les plus extrêmes, les cinq plus grandes et les cinq plus petites, ainsi que le numéro du cas correspondant à ces valeurs, tel qu'illustré dans le tableau ci-bas.

Extreme Values

			Case Number	Value
Age When First Married	Highest	1	1241	58
		2	190	54
		3	822	50
		4	744	49
		5	777	47
	Lowest	1	1357	13
		2	1377	14
		3	893	14
		4	763	14
		5	665	14

On voit que le cas numéro 1241 s'est mariée pour la première fois à 58 ans, et que c'est le cas numéro 1357 qui s'est mariée à 13 ans. Ceci nous permet d'examiner les autres caractéristiques de ces cas extrêmes, et qui nous apprend entre autres que les deux sont des femmes, et que celle qui s'est mariée à 13 ans avait au moment de l'enquête 67 ans.

L'option **Percentiles** produit un tableau de certains des percentiles, tel qu'illustré ci-bas. Les 25^e, 50^e, et 75^e percentiles sont appelés **Tukey's Hinges**. Ce sont ces valeurs qui sont utilisées pour constituer la partie centrale du graphique des « boîtes et moustaches », appelées aussi diagrammes en boîtes. (boxplots en anglais).

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	Age When First Married	17,00	18,00	19,00	22,00	25,00	29,00	32,00
Tukey's Hinges	Age When First Married			19,00	22,00	25,00		

Les options du bouton Plots

Le bouton **Plots** permet de déterminer quels graphiques on souhaite obtenir (figure 4.4). Par défaut, on obtient un diagramme « boîte et moustaches » (ou encore diagramme en boîtes), Boxplots, qu'on peut supprimer. On peut aussi obtenir les diagrammes descriptifs classiques, soit des diagrammes de dénombrement des cas (Stem-and-leaf) ou des histogrammes. Quant à l'option Normality plots with tests elle se rapporte à l'inférence statistique et ne sera pas traitée dans ce cours.

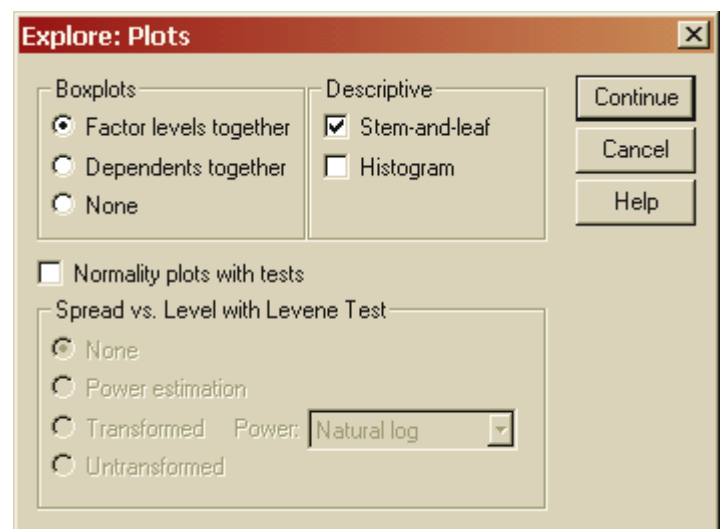


Figure 4.4

La ventilation des données en fonction de sous-groupes

Un des avantages de la commande **Explore...** est qu'elle permet d'obtenir les mesures descriptives qu'on a vues ci-haut séparément pour des groupes distincts, définis par une variable qualitative. Par exemple, on pourrait obtenir l'âge au premier mariage séparément pour les hommes et les femmes, ou séparément pour les divers niveaux d'éducation. Et là il n'est pas nécessaire que cette variable qualitative soit dichotomique et elle peut comporter plusieurs catégories. Nous allons illustrer cette procédure en l'appliquant à la variable de l'âge au premier mariage, ventilée en fonction du sexe. Voici les étapes pour le faire.

1. Dans la boîte de dialogue de la commande **Explore...**, placez la variable **Age When First Married** dans l'espace désigné pour la **Dependent List**, et placez la variable **Respondent's Sex** dans l'espace désigné par le terme **Factor List**.
2. Cliquez **OK** ou utilisez la syntaxe si vous préférez.
3. Vous obtiendrez toutes les statistiques usuelles de la commande **Explore...** séparément pour les hommes et pour les femmes. Les diagrammes en boîtes pour les hommes et pour les femmes seront juxtaposés, permettant des comparaisons.

IMPORTANT : Les variables placées dans la boîte **Dependent list** doivent obligatoirement être quantitatives. Les variables placées dans la boîte **Factor list** doivent obligatoirement être organisées en un petit nombre de catégories.

Exercice 4.1

1. Faites l'analyse de l'âge au premier mariage en fonction de la variable **degree2**. Écrivez vos conclusions en phrases complètes.
2. Refaites-la en fonction de l'appartenance religieuse.

Labo 5 La manipulation des données et des variables

Alors que les analyses statistiques proprement dites se font à travers le menu **Analyze**, la manipulation des données et des variables se font à travers les menus **Data** et **Transform**.

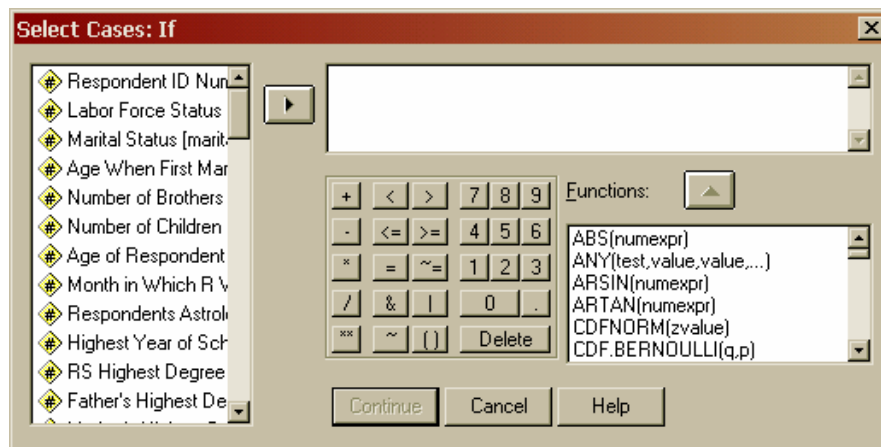
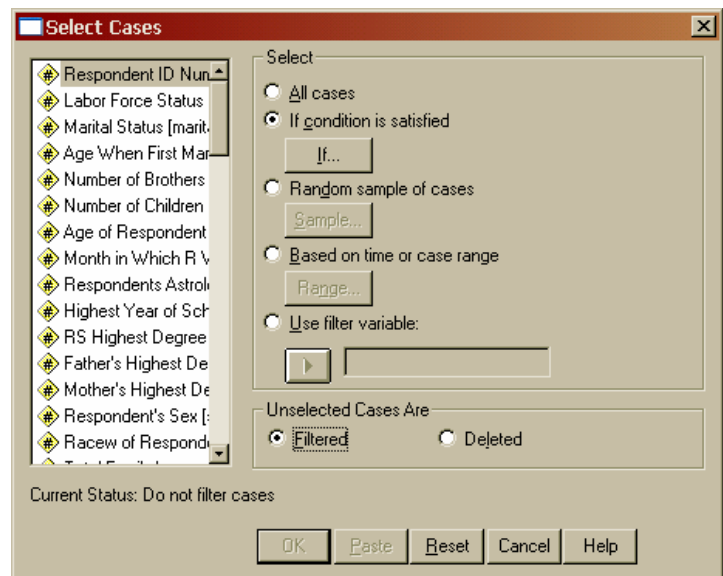
Nous allons faire quelques exercices comportant ces manipulations.

1. Sélectionner un sous-ensemble des données pour en faire un traitement statistique

Il y a deux façons de sélectionner un sous-ensemble de données : soit en effaçant les données qu'on ne souhaite pas conserver, ou alors en les conservant dans le fichier sans en tenir compte dans l'analyse. Ceci s'appelle 'filtrer' les données. Voici comment cela est effectué.

Dans le menu Data, cliquez sur **Select Cases**. Vous obtenez la fenêtre ci-contre. Vous remarquerez que le bouton **Filtered**, au bas de la fenêtre, est sélectionné. Toutes les données seront donc conservées dans le fichier. Nous avons cliqué sur l'option **If condition is satisfied**.

Il faut maintenant cliquer sur le bouton **If...** afin d'indiquer comment le filtrage de données doit se faire. En cliquant dessus, on obtient la boîte de dialogue suivante :



Supposons qu'on veuille choisir les hommes de cette population. Mais nous ne souvenons plus si les hommes sont codés 1 ou 2. Alors on clique sur la fenêtre **Variables** dans le menu **Utilitaires**. Mais SPSS refuse de réagir, car la boîte de dialogue de la commande **Select** est encore ouverte. Il faut la fermer, puis retourner à la commande **Variables**. On fait alors défiler les variables pour faire apparaître la variable **Respondent's Sex**. On obtient ce qui suit.

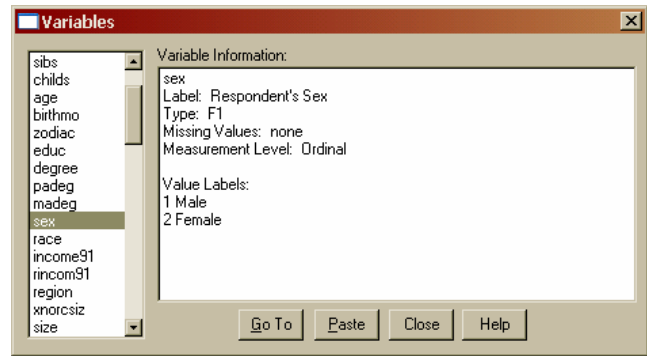
On voit que cette variable est codée ainsi : Hommes, codés 1; femmes, codées 2.

On retourne à la fenêtre **Select Cases**, on clique le bouton **If...** et on inscrit ce qui suit dans l'espace en haut à droite de la boîte de dialogue:

Sex = 1

On clique Continue, puis **OK** (rappelez-vous qu'on aurait pu cliquer Paste, ce qui nous aurait permis de conserver la syntaxe de cette commande).

On observe alors que dans la fenêtre des données, tous les numéros des cas non retenus (ici, il s'agit des femmes) sont barrés. Les analyses qui suivent vont se faire sans inclure ces cas.



Si on avait choisi **Delete** dans la première boîte de dialogue (**Select Cases**) les cas non retenus auraient été effacés du fichier. Dans ce cas, sauvegarder le fichier équivaut à effacer les cas non retenus pour de bon !! Si on veut vraiment travailler avec une partie des données seulement, il vaut mieux sauvegarder le fichier en changeant son nom. Ainsi, le fichier original sera conservé avec toutes les données, et le fichier de données modifié sera conservé sous un autre nom. Si on a choisi **Delete** par erreur et qu'on ne veut pas perdre les données, il faut fermer le fichier de données **SANS LE SAUVEGARDER** (donc en répondant **NO** à la question : **Save contents of data editor to nom_du_fichier ?**

(On pourrait aussi sauvegarder le fichier en changeant son nom : ainsi le fichier original resterait intact.

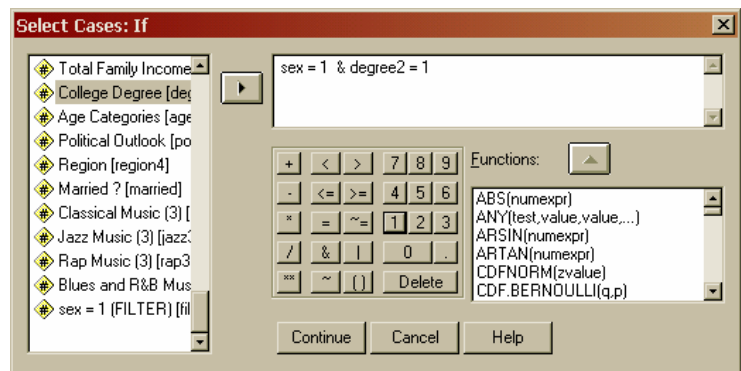
On peut toutefois sauvegarder le fichier de syntaxe et celui des tableaux et des graphiques, le **Output** qui est de type Viewer).

Exercice 5.1

Nous travaillons avec le fichier **GSS93 subset**. Sélectionnez les hommes de la population, puis faites un tableau des fréquences de leur statut d'emploi (par le biais de la syntaxe comme on l'a appris précédemment). Copiez ce tableau dans un document Word et ajoutez une explication de ce que vous avez fait. Ensuite, revenez à la commande **Select Cases**, cliquez l'option **All Cases** puis cliquez **OK**. Ensuite, produisez le tableau des fréquences des statuts d'emploi avec la même syntaxe. Assurez-vous que ce tableau comptabilise bien TOUTES les données du fichier.

Exercice 5.2 : Comment utiliser deux critères dans la sélection

Si vous voulez sélectionner les hommes qui ont un diplôme universitaire, refaites les étapes suivantes, mais quand vous spécifiez le critère de sélection (le bouton If...) il faut entrer les deux conditions, soit en les dactylographiant, soit en manipulant les variables à l'aide de la souris. La boîte de dialogue ressemblera alors à ce qui suit. Si vous cliquez Continue puis Paste, la syntaxe de la commande sera :



USE ALL.

COMPUTE filter_\$=(sex = 1 & degree2 = 1).

VARIABLE LABEL filter_\$ 'sex = 1 & degree2 = 1 (FILTER)'.
 VALUE LABELS filter_\$ 0 'Not Selected' 1 'Selected'.
 FORMAT filter_\$ (f1.0).
 FILTER BY filter_\$.
 EXECUTE .

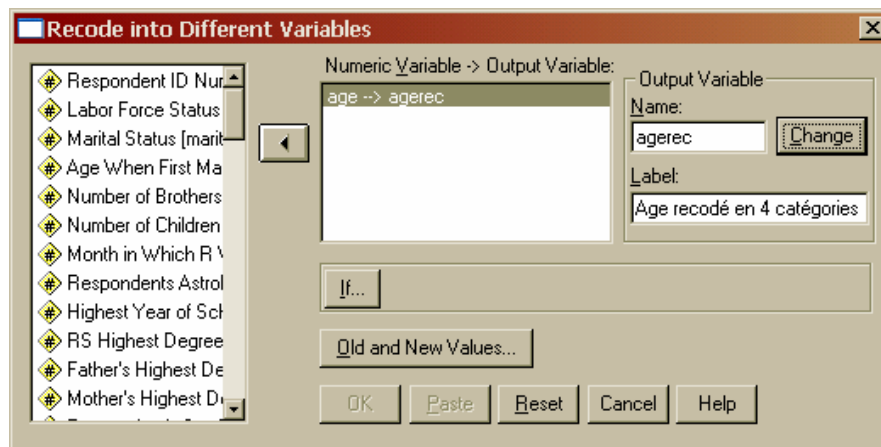
Produisez à présent les tableaux de fréquences produits précédemment pour cette sous-population.

2. Recodage d'une variable

Recoder une variable, c'est modifier la façon dont les valeurs et les catégories sont notées. En général, cette opération est entreprise pour regrouper des catégories distinctes. Pare exemple, si nous avons l'âge des répondants en année, nous pourrions vouloir les regrouper en tranches d'âge, en fonction d'une problématique qui nécessite un tel regroupement.

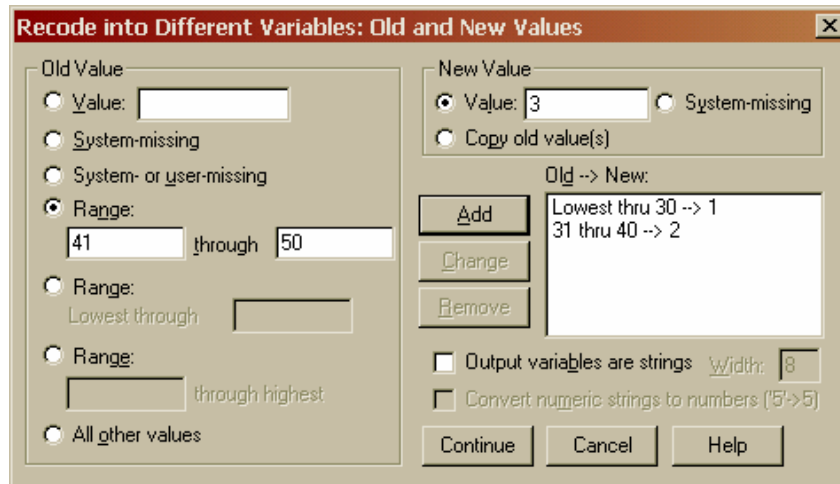
Le recodage dans SPSS se fait par la commande **Recode** qui se trouve dans le menu **Transform**. Quand on la sélectionne, on a le choix entre remplacer la variable qu'on recode par celle qui en résulte ('**Into same variable**') ou encore en créer une nouvelle qui résulte du recodage. Il vaut mieux toujours en créer une nouvelle car on pourrait vouloir réutiliser les valeurs originales. Nous allons effectuer l'opération suivante : créer une nouvelle variable par recodage qui comporte les catégories d'âge suivantes : 30 ans ou moins, 31 ans à 40 ans, 41 ans à 50 ans, puis 51 ans ou plus. Ceci se fait ainsi.

1. Sélectionnez **Recode → Into Different Variables...** sous le menu **Transform**.
2. Dans la boîte de dialogue qui apparaît, place la variable age dans l'espace prévu à droite à l'aide du bouton comportant une flèche en forme de triangle.
3. À droite, dans l'espace intitulé Name, inscrivez le nom de la nouvelle variable : agerec.
4. Inscrivez l'étiquette (Label) de cette variable, c'est-à-dire son nom tout au long : Age recodé en 4 catégories.
5. Cliquez sur le bouton **Change**. Le nom de la nouvelle variable va s'inscrire dans la fenêtre du centre, tel qu'illustré ci-bas.

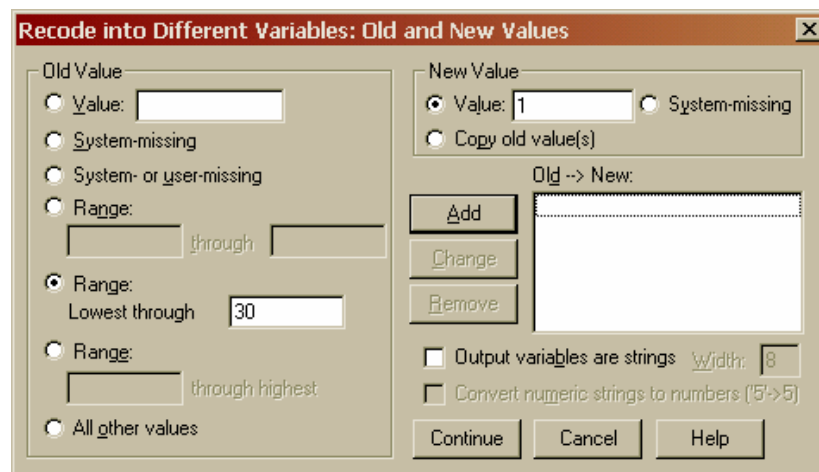


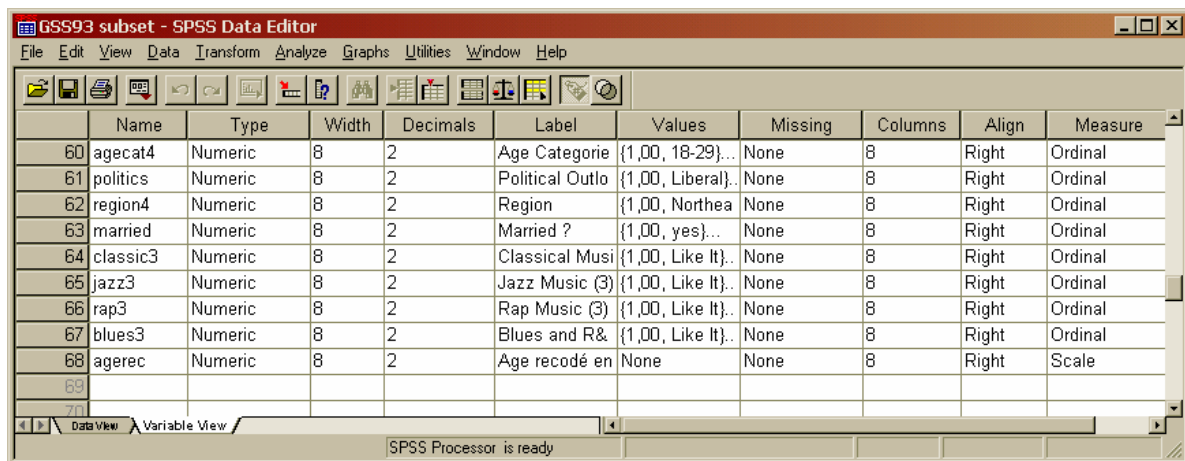
6. La nouvelle variable est créée, mais il faut indiquer comment définir les nouveaux codes, et ensuite donner des noms aux catégories ainsi définies. Pour définir les nouveaux codes, on clique sur le bouton **Old and New Values...**. Vous obtenez la fenêtre illustrée au haut de la page suivante. À gauche, on indique les anciennes valeurs, et on a plusieurs choix possible (une valeur en particulier, les valeurs comprises entre deux nombres, les valeurs plus grandes ou plus petites qu'un nombre, etc). À droite, on inscrit la nouvelle valeur qui remplace les anciennes. Ainsi, si on veut remplacer tous les ages de 30 ans ou moins par le code 1, on clique : **Range : Lowest through :**, on inscrit **30** dans l'espace approprié, on inscrit **1** dans l'espace intitulé **Value :**. La fenêtre de la page suivante inclut toutes ces opérations.

7. À présent il faut cliquer le bouton **Add** pour que ce changement soit enregistré dans la liste de calcul des nouvelles valeurs. On refait la même chose avec les valeurs 31 à 40 ans et 41 et 50 ans, puis 51 ans et plus, en prenant soin de choisir le bouton approprié à gauche : le bouton des valeurs entre deux nombres est différent de celui des valeurs plus petites qu'un certain nombre. Voici à quoi ressemble la fenêtre en cours d'opération.

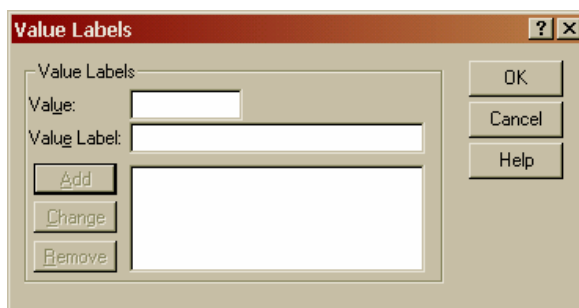


8. Quand on a rentré les quatre nouvelles catégories, on clique **Continue**. On retombe sur la première boîte de dialogue de la commande Recode. Plutôt que cliquer **OK**, cliquez **Paste** pour voir à quoi ressemble la syntaxe de cette commande. Vous devriez obtenir la syntaxe suivante.
- ```
RECODE
 age
 (Lowest thru 30=1) (31 thru 40=2) (41 thru 50=3) (51 thru Highest=4) INTO agerec .
VARIABLE LABELS agerec 'Age recodé en 4 catégories'.
EXECUTE .
```
9. Observez bien les termes entre parenthèses : ils indiquent comment on regroupe les anciennes valeurs et par quoi il faut les remplacer. Cette commande donne aussi le nom de la nouvelle variable (INTO agerec) et la ligne suivante, la commande VARIABLE LABELS, donne aussi le nom complet ou étiquette de la nouvelle variable. Si vous sélectionnez cette syntaxe et la faites exécuter, SPSS créera une nouvelle colonne à l'extrémité droite du fichier de données qui comportera ces nouvelles valeurs : 1, 2, 3, ou 4.
10. Ces nouvelles catégories n'ont pas encore de noms ! Il faut en mettre. Pour cela, aller à l'affichage des variables, du fichier des données, et faites défiler la liste des variables vers le bas. La dernière variable est celle que nous venons de créer. La fenêtre devrait ressembler à ceci :





11. Cliquez sur la colonne Decimals dans la case correspondant à agerec, et changez le 2 pour un 0 : nous n'avons pas besoin de décimales pour noter ces quatre catégories.
12. Cliquez sur le côté droit de la case Values correspondant à la variable agerec. En cliquant deux fois, la boîte de dialogue suivante devrait apparaître :



13. Cette boîte de dialogue va vous permettre d'inscrire les quatre catégories créées, une à une, et de leur coller une étiquette (Value Label). Ainsi vous auriez :  
**Value : 1**  
**Value Label : Moins de 30 ans.**  
 Ensuite vous cliquez Add, et vous faites de même pour les trois autres catégories. Cliquez **OK**, et aller vérifier dans le fichier de données. Vous verrez que les catégories on à présent une étiquette, et ce sont ces étiquettes qui vont apparaître quand vous produisez des tableaux.
14. Produisez le tableau de fréquence de la nouvelle variable pour vous assurer que les catégories ont bien été créées correctement.
15. Sauvegarder le fichier de données sous un nouveau nom (ex : GSS93 recodé) sur une disquette ou dans votre dossier personnel si vous en avez un sur le serveur.

### Exercice 5.3

Recodez la variable 'age au premier mariage' en 5 catégories ainsi :

Moins de 18 ans, de 18 à 25 ans, de 26 à 35 ans, de 36 à 45 ans, plus de 45 ans.

Produisez le tableau de fréquences de la nouvelle variable ainsi créée.

## Labo 6 La création de nouvelles variables à l'aide de la commande Compute

SPSS nous permet de créer de nouvelles variables à partir de variables existantes, en utilisant la commande Compute. Cette commande permet de définir cette nouvelle variable et de calculer les valeurs qu'elle prend pour chacun des cas. Pour effectuer ce calcul, nous pouvons nous-mêmes spécifier les opérations arithmétiques à faire (additionner ou retrancher des valeurs existantes, les multiplier ou les diviser, etc.) ou encore utiliser des fonctions déjà inscrites dans SPSS, telles que : mettre une valeur au carré, calculer le logarithme d'une valeur (pouah !), calculer la racine carrée d'une valeur, choisir la plus grande de deux valeurs ou même de n valeurs, coller ensemble deux suites de chiffres ou de lettres, sélectionner les 3 premiers chiffres (ou les n premiers) d'une suite de lettres, calculer la différence entre deux dates, etc. Par exemple, si nous avons une variable qui donne l'âge du répondant, et une autre qui donne la durée de son mariage ou de son union civile, nous pouvons calculer l'âge lors du mariage ainsi :

$$\text{Âge du répondant lors du mariage} = \text{âge actuel} - \text{durée du mariage.}$$

SPSS nous permet même de faire des calculs logiques qui incluent des conditions reliées par ET ou par OU. Par exemple, si on a les trois variables :

Vit seule (oui/non)  
A des enfants (oui/non)  
Travaille (oui/non)

on peut inventer une variable qu'on appellerait 'solitude' et qu'on définirait ainsi :

Si la personne vit seule ET qu'elle n'a pas d'enfant ET qu'elle ne travaille pas : solitude = 4  
Si la personne vit seule ET qu'elle a des enfants ET qu'elle ne travaille pas : solitude = 3  
Si la personne vit seule ET qu'elle a des enfants ET qu'elle travaille : solitude = 2  
Si la personne ne vit pas seule : solitude = 1

Il faudrait aussi qu'on ait une bonne raison de définir la solitude ainsi et que cette définition ait un sens dans le cadre théorique auquel on se réfère. (Vous aurez remarqué que la définition donnée ci-haut est incomplète, car elle n'assigne pas de valeur à la variable 'solitude' dans un des cas possibles : essayez d'identifier ce cas et proposez une valeur qui vous semble raisonnable).

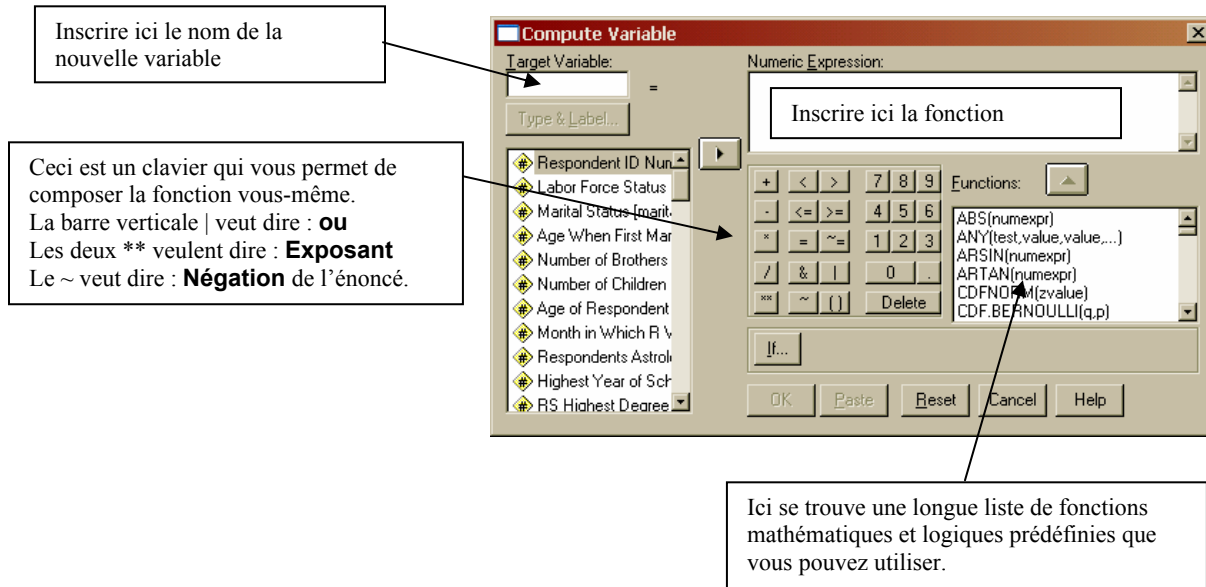
Nous allons examiner à présent comment utiliser la fonction Compute à l'aide des menus ainsi qu'à l'aide de la syntaxe. Ouvrez le fichier **GSS93**. Nous souhaitons calculer une fonction appelée **Durée du mariage** que nous voulons calculer ainsi :

$$\text{Durée du mariage} = \text{âge actuel} - \text{âge au premier mariage.}$$

Mais il y a un problème avec cette définition. Qu'est-ce qui garantit que ce premier mariage a perduré jusqu'au moment de l'enquête ? Nous n'avons pas de variable qui nous informe si c'est toujours le premier mariage qui est en cours. On ne peut pas non plus utiliser la variable Statut civil (Marital Status) pour tenir compte des personnes veuves, car même si la personne a répondu qu'elle est mariée, nous ne savons pas si ceci est le premier mariage.... Compte tenu des informations que nous avons, tout ce qu'on peut calculer, c'est une variable qu'on appellerait : temps écoulé depuis le premier mariage, qu'on obtiendrait en faisant la différence entre l'âge actuel et l'âge au premier mariage. Donc, on ne suppose pas que c'est le premier mariage qui est en cours. Pour simplifier les choses, nous allons calculer la variable : Année de la naissance. Comme l'enquête a été effectuée en 1993, il suffira de retrancher l'âge de la personne de 1993. La formule sera donc :

### Année de la naissance = 1993 – (âge actuel)

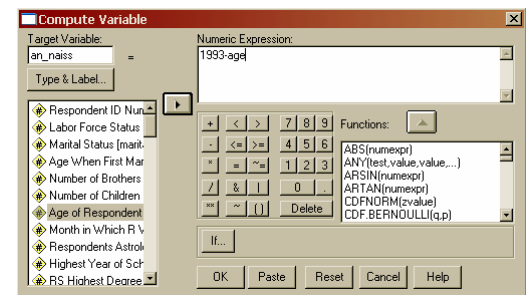
Voici comment cela est fait dans SPSS. Cliquer sur **Transform → Compute**. Vous obtenez la fenêtre suivante.



Nous allons appeler notre nouvelle fonction : **an\_naiss**, et son nom au complet ou étiquette (le *Variable Label*) sera : **Année de naissance**. Inscrivez : **an\_naiss** dans l'espace approprié, puis inscrivez dans l'espace intitulé **Numeric Expression** la formule suivante :

**1993 – age**

Vous n'allez pas écrire le mot **age**. Vous allez plutôt cliquer sur la variable **age** dans la liste qui se trouve à gauche, et puis cliquer sur le petit triangle noir qui va placer la variable **age** là où se trouve le curseur. Votre boîte de dialogue ressemblera alors à ceci : Cliquez maintenant sur le bouton **Type & Label**, et inscrivez le nom de la variable au complet : **Année de naissance**. Assurez-vous que la variable est de type numérique, puis cliquez **Continue**. Vous revenez à la boîte de dialogue intitulée **Compute Variable** illustrés ci-haut.



Si vous cliquez **OK**, la nouvelle variable sera créée et se retrouvera au bout de votre fichier (à droite dans l'affichage des données, et en bas dans l'affichage des variables). Mais plutôt que de cliquer **OK**, nous allons cliquer **Paste**. La commande sera alors affichée ainsi dans la fenêtre de syntaxe :

```
COMPUTE an_naiss = 1993-age .
VARIABLE LABELS an_naiss 'Année de naissance' .
EXECUTE .
```

Regardez bien chaque ligne. La première définit la nouvelle variable, et la deuxième spécifie le nom au complet (l'étiquette) de cette nouvelle variable. Sélectionner la syntaxe et faites-la exécuter, puis courez voir dans la fenêtre des données si elle a été créée comme il faut.



**Exercice 6.1**

Créez la variable **Année du premier mariage** et produisez un histogramme de fréquences. Copiez l’histogramme dans un document Word, copiez aussi la syntaxe qui a produit cette variable, et expliquez pourquoi il y a des données manquantes. **Conservez ce travail dans vos dossiers.**

## Labo 7: Création d'un fichier de données

Le but de cette leçon est d'apprendre comment créer un fichier de données à partir d'un questionnaire, comment inclure toutes les caractéristiques exigées des variables qui sont créées, comment saisir des données, et comment sauver et imprimer un fichier de données et un fichier de résultats.

### 1. Ouverture d'un nouveau fichier de données

Rappelez-vous qu'un fichier de données électroniques se compose de l'information qui a été rassemblée, puis saisie à l'aide d'un logiciel statistique et organisée de manière à faciliter son analyse statistique. Chaque colonne représente une **variable**, et chaque ligne représente un **cas**.

Quand vous ouvrez SPSS, vous obtenez une fenêtre qui offre plusieurs options, incluant:

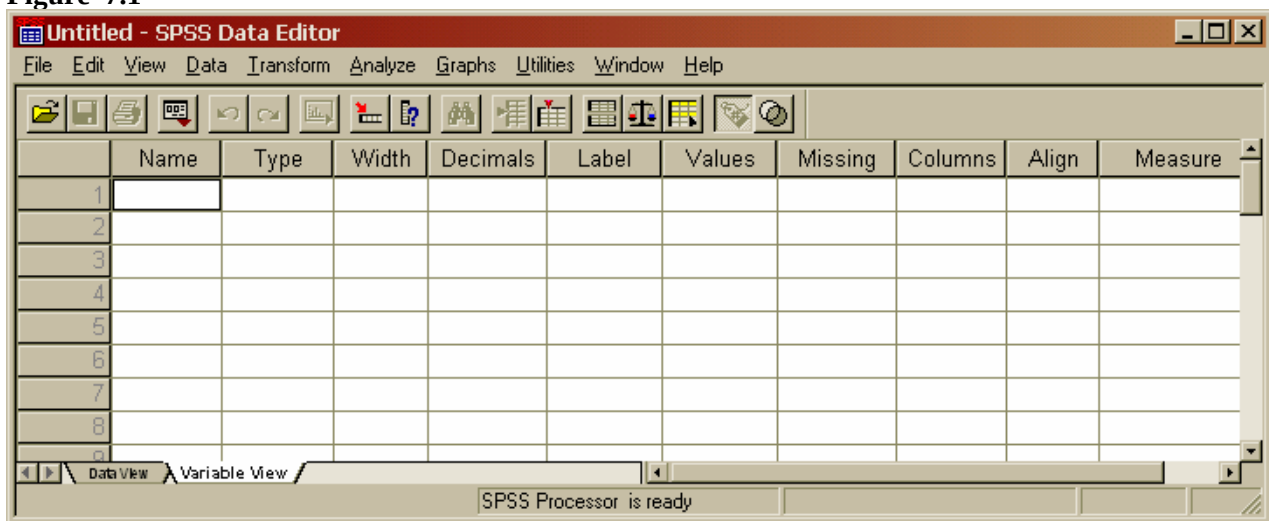
**O Type in data (Saisir les données)**

Et

**O Open a data file (Ouvrir un fichier de données existant).**

Cliquez le bouton rond précédant les mots **Type in data**: vous obtenez un fichier de données vierge. Parfois, le programme est configuré de telle sorte qu'il s'ouvre en mode **Saisie des données**. Vous obtenez la fenêtre illustrée dans la figure 7.1

Figure 7.1



Les colonnes correspondent aux variables, et les lignes aux cas. Avant de saisir les données, vous devez indiquer les caractéristiques de chacune des variables que vous voulez utiliser. Il est toujours préférable de préparer une matrice de données vide avant de saisir les données, et d'imprimer l'information du fichier pour vérifier si les variables ont été définies correctement. Nous pouvons toujours modifier les caractéristiques d'une variable par la suite, et ajouter même de nouvelles variables, mais il est préférable de démarrer avec une bonne matrice vierge où toutes les variables ont été définies correctement. Le mot « matrice » est employé pour indiquer un fichier de données vierge où les variables ont été définies. Une fois que les données sont saisies, nous désignerons le fichier qui en résulte comme étant un fichier de données.

Pour construire la matrice de SPSS, cliquez sur l'onglet **Variable View** qui apparaît au bas de la fenêtre du **Data Editor**.

Le nom du fichier apparaît au haut de la fenêtre. Dans cet exemple, le mot **Untitled** apparaît parce que nous n'avons pas encore donné un nom au fichier de données.

Examinez soigneusement les diverses colonnes de l'affichage **Variable View**. Chaque ligne est une variable, et chaque colonne permet de déterminer l'une des caractéristiques des variables à définir.

Nous allons illustrer dans ce qui suit comment remplir ces cases, mais indiquons en attendant ce que les diverses colonnes permettent de spécifier :

**Name** : c'est le nom de la variable (8 lettres ou chiffres, sans espace)

**Type** : c'est son type (des nombres, ou une date, ou des lettres etc. )

**Width** : le nombre d'espaces réservés pour inscrire les valeurs

**Decimals** : c'est le nombre de décimales utilisées pour cette variable

**Label** : c'est le nom complet de la variable, qui apparaîtra dans les tableaux

**Values** : on inscrit ici les catégories désignées par les codes utilisés pour mesurer les valeurs

**Missing** : on indique ici les valeurs qui doivent être considérées manquantes

**Columns** : on indique ici la largeur des colonnes souhaitée pour l'affichage Data

**Align** : permet d'aligner les valeurs de cette variable à gauche, à droite ou au centre

**Measure** : permet d'indiquer quelle échelle de mesure est utilisée (nominale, ordinale ou numérique).

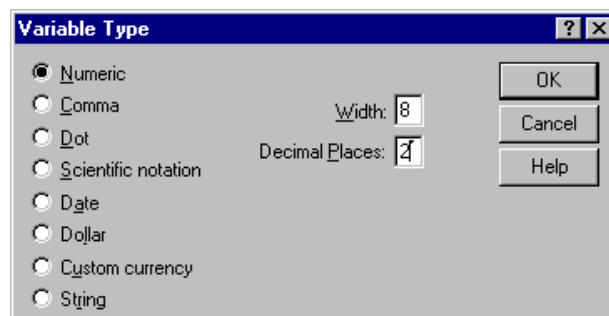
Nous illustrerons le procédé avec la variable **Sexe du répondant**. Nous utiliserons le code suivant :

1 Homme

2 Femme,

et nous lui attribuons le nom '**sexe**'.

1. La première colonne à compléter est le **nom** de la variable. Celui-ci doit être un nom court, avec tout au plus 8 caractères et aucun espace. Dactylographiez le nom 'sexe'.
2. La deuxième colonne est le **Type**. Cliquez sur le côté droit de la cellule ; vous obtenez la zone de dialogue montrée dans fig. 7.2. Puis, vous indiquez si les données sont numériques, ou un signe tel qu'un point ou une virgule, ou une devise, ou une date, ou une variable chaînes de caractères.



**Figure 7.2**

Une variable de type **Numérique** est employée quand vous voulez saisir vos données à l'aide de codes numériques, comme la variable **Sexe** : les codes employés pour la saisir (1 ou 2) sont des nombres. Si ces nombres se rapportent à des valeurs numériques réelles vous devez indiquer le nombre d'espaces et de décimales dont vous avez besoin. Par exemple, si vous voulez enregistrer la taille du répondant mesurée en centimètres, vous avez besoin de 5 chiffres avec une seule décimale, afin de pouvoir écrire des nombres comme **172.3** centimètres (le point emploie un espace). Si les nombres se rapportent à des catégories (par exemple : 1 = Homme ; 2 = Femme), vous n'avez besoin que d'un seul espace sans aucune décimale. Vous aurez besoin de deux espaces si vous avez plus de dix mais moins de 100 catégories, et ainsi de suite.

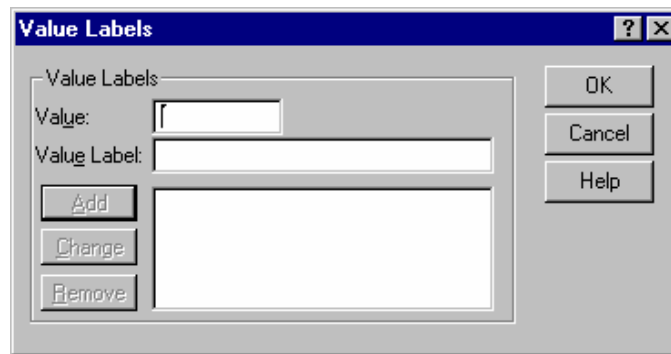
Une variable de type **String** comporte des valeurs qui sont des suites de lettres ou de chiffres sans valeur numérique. Elle est employée quand vous voulez saisir un nom propre par exemple, tel que Pierre, ou Marie. Peu de procédures statistiques s'appliquent aux variables de type **String**. Elles sont utilisées pour désigner les divers cas du fichier, ou encore pour retranscrire des questions ouvertes.

3. La troisième colonne, **width**, permet de spécifier le nombre d'espaces requis est largeur. On a déjà spécifié ce nombre à l'étape précédente, mais il peut être modifié directement dans cette colonne.

4. Les mêmes remarques s'appliquent à la quatrième colonne qui permet de spécifier le nombre de décimales.

5. La cinquième colonne, **Label**, est le nom détaillé de la variable, par exemple : **Sexe du répondant**. C'est le nom qui apparaîtra dans les tableaux produits par SPSS. Il faut donc le choisir avec soin pour qu'il désigne clairement la variable tout en étant concis.

6. La sixième colonne permet de spécifier les catégories utilisées pour mesurer la variable. Cliquez sur le côté droit de la cellule. La zone de dialogue montrée dans la figure. 7.3 devrait apparaître.



**Figure 7.3**

Inscrivez

1 dans la case **Value**,  
et  
Homme dans la case **Value Label**.

Cliquez sur **Add**, puis recommencer avec :

2 dans la case **Value**,  
et  
Femme dans la case **Value Label**.

Cliquer encore sur **Add**, puis sur **OK**.

7. La septième colonne permet de spécifier les valeurs manquantes. Pour la variable **Sexe**, nous pourrions la laisser telle quelle puisqu'on s'attend à ce que le sexe du répondant soit connu. Mais nous pouvons penser aux situations où le sexe du répondant n'est pas connu avec certitude (par exemple s'il est déterminé par la voix dans une communication téléphonique) ou si les données proviennent de fichiers d'archives incomplets. Dans ces cas, il faut prévoir une valeur manquante. En cliquant sur le côté droit de la cellule, nous obtenons la zone de dialogue montrée dans la figure 7.4.

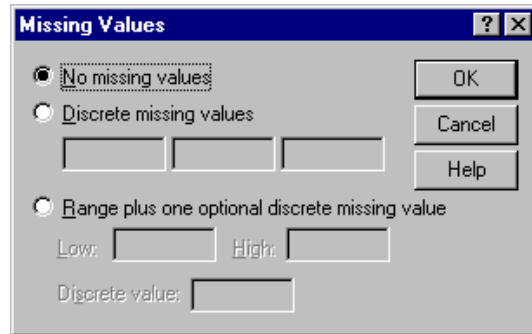


Figure 7.4

La fenêtre vous donne trois choix. Ou vous n'avez aucune valeur manquante, ou vous avez jusqu'à trois valeurs manquantes distinctes (c'est ce que signifie le mot **Discrete**) ou enfin vous considérez comme manquantes toutes les valeurs qui tombent dans l'étendue entre deux nombres, avec la possibilité d'avoir une valeur manquante distincte additionnelle. Par exemple, on peut avoir les valeurs manquantes codées par : 7, 8, et 9. Mais il faudra retourner à la colonne **Values** et inscrire les significations des codes 7, 8 et 9. Par exemple, cela pourrait être :

- 7 Ne sait pas
- 8 Pas de réponse, et
- 9 Ne s'applique pas.

Par exemple, si vous aviez deviez coder la question posée uniquement aux personnes qui remplissent une déclaration de revenus :

*Pensez-vous que les politiques fiscales du gouvernement sont bonnes ?*

vous pourriez utiliser les codes suivants et distinguer les données manquantes par les trois possibilités : **Ne sait pas**, **Refuse de répondre**, ou **Ne s'applique pas**. La possibilité **Ne s'applique pas** serait cochée pour les personnes qui ne remplissent pas de déclaration de revenu. Ceci est illustré par la figure 7.5.

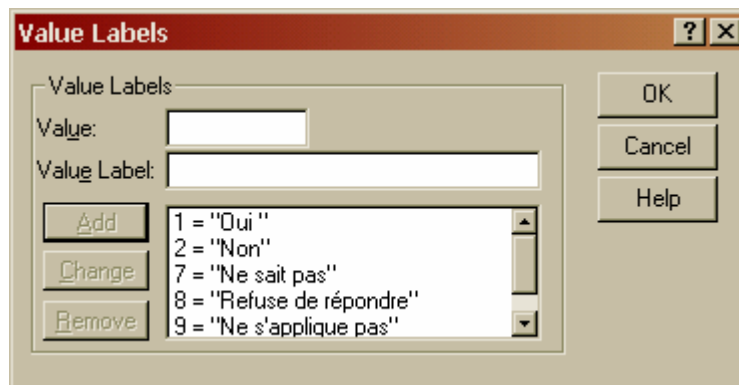


Figure 7.5

Notez qu'il est important d'indiquer les valeurs manquantes correctement, sans quoi SPSS les comptera dans les statistiques, ce qui pourrait fausser les résultats.

8. La huitième colonne, **Columns**, vous permet de déterminer la largeur de l'affichage de la colonne à l'écran. Vous pouvez laisser la valeur suggérée par SPSS par défaut.

9. La neuvième colonne, **Align**, vous permet de déterminer l'alignement (gauche, droite ou centre) du texte apparaissant à l'écran dans l'affichage **Data View**. Vous pouvez laisser la valeur suggérée par SPSS par défaut.

10. La dernière colonne, **Mesure**, est très importante. Elle permet de déterminer l'échelle de mesure utilisée pour la variable : numérique, ordinale ou nominale (**scale, ordinal, nominal**). Pour la variable **sexe**, nous choisirons **nominal**. Vous avez peut-être noté que certains des fichiers d'exemples de SPSS se

classent des variables qualitatives telles que **sexe** comme étant ordinales. Ceci est généralement fait quand une variable qualitative a seulement deux catégories (c'est-à-dire quand elle est dichotomique), et ceci nous permet d'exécuter certaines procédures statistiques sur de telles variables. Dans ce cours, nous n'aurons pas besoin d'employer les échelles ordinales de mesure pour coder des variables qualitatives.

### Vérification de la matrice SPSS créée

Quand vous aurez fini de définir toutes vos variables, vous pourriez vouloir vérifier que vous n'avez fait aucune erreur. La meilleure manière de le faire est de faire apparaître toutes les variables avec leurs caractéristiques et d'examiner les résultats. Ceci est fait en cliquant sur **Utilities→File Info**. Le résultat devrait ressembler à ce qui suit (nous avons défini la variable **Sexe**, et la variable **pol\_fisc** concernant les politiques fiscales mentionnées plus haut.

| List of variables on the working file |                                                                                                                                                                                                                                                                                                              |          |
|---------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|
| Name                                  |                                                                                                                                                                                                                                                                                                              | Position |
| SEXE                                  | Sexe du répondant<br>Measurement Level: Nominal<br>Column Width: 8 Alignment: Right<br>Print Format: F1<br>Write Format: F1<br><br>Value      Label<br><br>1      Homme<br>2      Femme                                                                                                                      | 1        |
| POL_FISC                              | Pensez-vous que les politiques fiscales du gouvernement sont<br>Measurement Level: Nominal<br>Column Width: 8 Alignment: Left<br>Print Format: F1<br>Write Format: F1<br><br>Value      Label<br><br>1      Oui<br>2      Non<br>7      Ne sait pas<br>8      Refuse de répondre<br>9      Ne s'applique pas | 2        |

Toutes les caractéristiques de la variable sont indiquées ici. Vous pouvez vérifier que le nom bref et le nom complet de chaque variable sont corrects. Remarquez que le libellé de la question sur les politiques fiscales est incomplet car il est trop long. Cependant, il apparaîtra au complet dans les tableaux de fréquences. Vérifiez que les codes sont bien ceux que vous vouliez saisir. Vous devez également examiner le format, les valeurs manquantes, ainsi que les codes attribués aux valeurs manquantes.

Quand vous avez défini toutes vos variables et que vous avez vérifié que cela a fait correctement, vous pourrez commencer à saisir vos données dans la fenêtre **Data View**. En général, il vaut . Après que vous

avez dactylographié les données dans une des cellules, si vous appuyez sur la toumieux le faire un cas à la fois, c'est-à-dire ligne par ligne. La touche **TAB** sur votre clavier vous permet de déplacer le curseur à la cellule suivante sur la même ligne, mais si vous appuyez sur la touche **Enter** sur votre clavier le curseur se déplacera à la cellule de la ligne suivante, dans la même colonne.

Quand vous saisissez les données, vous devez écrire les codes, et non pas les noms des catégories. Par exemple, pour la variable **sexe**, vous écrirez :

- 1 et non pas « Homme », ou
- 2 et non pas « Femme ».

SPSS fera apparaître soit les valeurs, soit les codes, selon que **Value Labels** est coché ou pas dans le menu **View**.

Pour une variable quantitative telle que l'âge du répondant, vous écrirez l'âge lui-même. Il n'y aura aucune « étiquette » de valeur pour cette variable, mais vous auriez intérêt à inclure une valeur manquante tel que 999 pour les cas où la réponse est manquante. Cependant, si l'âge est codé en catégories, alors là il faut saisir la désignation de chaque catégorie. Par exemple, l'âge peut être codé ainsi :

- 1 Moins de 25 ans
- 2 De 25 à 39 ans
- 3 De 40 à 64 ans
- 4 65 ans ou plus
- 999 Non réponse.

Mais il est préférable de noter l'âge en années ou en mois, et ensuite de regrouper en catégories si nécessaire.

## Exercice pratique

Créez un fichier de données de SPSS pour saisir des données recueillies à l'aide du questionnaire suivant. Un questionnaire non rempli est donné, suivi des données (hypothétiques) se rapportant à 10 enfants. Créez d'abord une matrice SPSS vierge incluant toutes les questions du questionnaire. N'oubliez pas d'inclure des valeurs manquantes chaque fois que cela est approprié. Puis imprimez l'information sur les variables (**File Info**) pour vérifier que les variables ont été créées correctement. Ensuite, vous pourrez saisir les données ci-dessous et enregistrer sur une disquette le fichier de données que vous aurez créé. Pour vous assurer d'avoir saisi les données correctement, produisez les tableaux de fréquence pour toutes les variables et examinez-le pour voir si ils correspondent aux données fournies.

Pour fin d'évaluation : recopiez l'information sur les variables (**File Info**) dans un document Word bien identifié (Labo 7, date, votre nom) et remettez-le.

**QUESTIONNAIRE**

Numéro du questionnaire: \_\_\_\_\_

1. Sexe : Garçon (M) \_\_\_\_\_ Fille (F) \_\_\_\_\_
2. Âge de l'enfant ? \_\_\_\_\_ (en mois)
3. Taille de l'enfant en centimètres ? \_\_\_\_\_
4. Taille du père (en cm) ? \_\_\_\_\_
5. Taille de la mère (en cm) ? \_\_\_\_\_
6. Couleur naturelle des cheveux ? (cochez une seule case)
  - ☐ Noirs
  - ☐ Bruns
  - ☐ Blonds
  - ☐ Roux
  - ☐ Châtains
7. Couleur des yeux ?
  - ☐ Noirs
  - ☐ Bruns
  - ☐ Bleus
  - ☐ Verts
  - ☐ Autre
8. Est-ce que l'enfant a déjà eu des accidents nécessitant une hospitalisation ? Oui \_\_\_\_\_ Non \_\_\_\_\_
9. Est-ce que l'enfant est gaucher, droitier, ou ambidextre? \_\_\_\_\_
10. Est-ce que l'enfant fréquente une garderie deux jours par semaine ou plus ? Oui \_\_\_\_\_ Non \_\_\_\_\_

**Voici les réponses concernant 10 enfants :**

| Q1 | Q2 | Q3 | Q4  | Q5  | Q6     | Q7    | Q8  | Q9     | Q10 |
|----|----|----|-----|-----|--------|-------|-----|--------|-----|
| M  | 20 | 68 | 172 | 170 | Noirs  | Verts | Non | Gauche | Non |
| F  | 18 | 67 | 180 | 165 | Bruns  | Bleus | Non | Droite | Oui |
| F  | 22 | 68 | 175 | 176 | Blonds | Noirs | Oui | Droite | Oui |
| F  | 22 | 67 |     | 169 | Noirs  | Bleus | Oui | Ambid. | Oui |
| M  | 20 | 68 | 176 | 164 | Bruns  | Bruns | Non | Droite | Non |
| M  | 28 | 76 | 171 | 166 | Roux   | Bleus | NSP | Droite | Oui |
| F  | 18 | 60 | 178 | 172 | Bruns  | Bruns | Oui | Droite | Non |
| M  | 17 | 61 | 177 | 167 | Bruns  | Bleus |     | Droite | Non |
| M  | 22 | 68 | 170 | 172 | Blonds | Bleus | Non | Droite | Non |
| F  | 30 | 78 | 168 | 160 | Bruns  | Bruns | Oui | Ambid. | Oui |



## Labo 8 Les tableaux croisés à deux entrées

On obtient les tableaux croisés par la commande **Analyze → Descriptive Statistics → Crosstabs**.

On place la variable indépendante dans la boîte des lignes (rows) et la variable dépendante dans celle des colonnes. On peut faire le contraire aussi : les résultats statistiques seront exactement les mêmes, mais les tableaux seront moins faciles à lire. Il est plus naturel de les lire quand on place les variables tel qu'il a été suggéré ci-haut.

Essayons avec les variables **College Degree** (qui se trouve vers la fin de la liste, dont le nom bref est **degree2**) comme variable dépendante, et **Respondent Sex** comme variable indépendante.

Pour obtenir les pourcentages de chaque catégories selon les lignes, on clique le bouton **Cells**, et dans la boîte des pourcentages, on sélectionne **Rows**.

Si on clique **Paste**, on obtient la syntaxe suivante :

```
CROSSTABS
 /TABLES=sex BY degree2
 /FORMAT=AVALUE TABLES
 /CELLS= COUNT ROW .
```

Rappelez-vous que vous pouvez aussi dactylographier cette syntaxe directement, sans passer par les menus. Si on exécute cette syntaxe, on obtient le tableau suivant.

| Respondent's Sex * College Degree Crosstabulation |          |                           |                     |                  |        |
|---------------------------------------------------|----------|---------------------------|---------------------|------------------|--------|
|                                                   |          |                           | College Degree      |                  | Total  |
|                                                   |          |                           | 0 No College degree | 1 College degree |        |
| Respondent's Sex                                  | 1 Male   | Count                     | 466                 | 175              | 641    |
|                                                   |          | % within Respondent's Sex | 72,7%               | 27,3%            | 100,0% |
|                                                   | 2 Female | Count                     | 683                 | 172              | 855    |
|                                                   |          | % within Respondent's Sex | 79,9%               | 20,1%            | 100,0% |
| Total                                             |          | Count                     | 1149                | 347              | 1496   |
|                                                   |          | % within Respondent's Sex | 76,8%               | 23,2%            | 100,0% |

Ce tableau nous donne le nombre de cas dans chaque cellule, mais aussi le pourcentage relatif à chaque ligne. Ainsi, on peut voir que chez les hommes, 72,7 % d'entre eux n'ont pas de diplôme universitaire, et que 27,3 % d'entre eux en ont obtenu un. Chez les femmes, seules 20,1 % ont un diplôme universitaire. Il y a donc une différence importante entre les hommes et les femmes de cet échantillon en ce qui concerne le taux de diplomation universitaire.

Mais attention : nous n'avons pas dit que cette relation est causale, ni qu'elle est valide pour l'ensemble de la population. Ces affirmations pourraient bien être vraies, mais les informations dont nous disposons ne nous permettent pas de le conclure.

Pour conclure savoir si cette relation est généralisable à la population entière, ou si elle est le fruit du hasard du choix de l'échantillon, il faut faire un test du Chi-carré. Les fondements de ce test seront discutés ultérieurement, mais on peut déjà apprendre à produire les mesures nécessaires et à les interpréter.

Pour que SPSS calcule le chi-carré, quand on donne la commande **Crosstabs**, on clique sur le bouton **Statistics** et on coche le choix **Chi-square**. On obtient la syntaxe suivante.

```

CROSSTABS
 /TABLES=sex BY degree2
 /FORMAT=AVALUE TABLES
 /STATISTIC=CHISQ
 /CELLS= COUNT ROW .

```

Remarquez la ligne qui se lit :

```
/STATISTIC=CHISQ
```

C'est une sous-commande qu'on peut ajouter ou non, et qui demande à SPSS de calculer le chi-carré. EN exécutant cette syntaxe, on obtient le tableau suivant :

| Chi-Square Tests                   |                     |    |                       |                      |                      |
|------------------------------------|---------------------|----|-----------------------|----------------------|----------------------|
|                                    | Value               | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square                 | 10,613 <sup>b</sup> | 1  | ,001                  |                      |                      |
| Continuity <sup>a</sup> Correction | 10,214              | 1  | ,001                  |                      |                      |
| Likelihood Ratio                   | 10,538              | 1  | ,001                  |                      |                      |
| Fisher's Exact Test                |                     |    |                       | ,001                 | ,001                 |
| Linear-by-Linear Association       | 10,606              | 1  | ,001                  |                      |                      |
| N of Valid Cases                   | 1496                |    |                       |                      |                      |

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 148,68.

Seule la première ligne nous intéresse. Elle indique que le chi-carré est égal à 10,613, une valeur que nous ne pouvons pas interpréter pour le moment. Mais elle indique aussi le niveau de signification dans la colonne **Asymp. Sig. (2-sided)**. Ce niveau est de 0,001. Cette mesure nous intéresse. Elle indique que :

**s'il n'y avait pas de différence entre les hommes et les femmes de la population, il y aurait moins de 1 chance sur 1000 que l'on obtienne un tel échantillon.**

Or ceci est si rare, que l'on préfère opter pour l'autre solution : celle de **supposer qu'il y a effectivement une différence entre les hommes et les femmes au niveau de toute la population.**

Les niveaux de signification peuvent être interprétés ainsi :

- Un niveau de 0,05 signifie qu'il y a moins de 5 pour 100 de chances d'obtenir un tel échantillon d'une population où il n'y aurait pas de différences entre les hommes et les femmes.
- Un niveau de 0,01 signifie qu'il y a moins de 1 pour 100 de chances d'obtenir un tel échantillon d'une population où il n'y aurait pas de différences entre les hommes et les femmes.

**Exercice 8.1:** (retranscrire les énoncés sur un document Word ou sur une feuille à remettre)

1. Produire le tableau croisé pour les variables **Sex** (indépendante) et **vote92** (dépendante), en produisant les pourcentages par ligne et le chi-carré. Répondre aux questions suivantes :

- a) Le pourcentage d'hommes de cet échantillon ayant voté en 92 est de .....
- b) Le pourcentage d'hommes de cet échantillon ayant voté en 92 est de .....
- c) Le pourcentage de personnes de cet échantillon ayant voté en 92 est de .....
- d) Les hommes de cet échantillon ont tendance à voter ..... (plus / moins) que les femmes.
- e) Le chi-carré est de ..... et le niveau de signification de .....
- f) Ceci signifie que la relation entre le sex et le fait de voter ou non est ..... (significative/non-significative).

2. Recommencer l'exercice avec les variables **GUNLAW** et **CAPPUN**.

## Labo 9 Les comparaisons de moyennes

Cette procédure est utilisée pour voir s'il y a un lien entre une variable qualitative (ou organisée en un nombre restreint de catégories), considérée comme la variable indépendante, et une variable quantitative considérée comme la variable dépendante. On calcule la moyenne de la variable quantitative sur chacune des catégories de la variable indépendante, pour voir s'il y a des différences notables. Ici, nous utiliserons aussi les diagrammes en boîtes (box-plots) pour illustrer ces différences. Par exemple, on peut calculer la moyenne du revenu pour les hommes et pour les femmes séparément.

Pour utiliser cette procédure, utiliser la commande **Analyze→Compare Means→Means**. Inscrivez une variable quantitative dans la case **Dépendent List** et une variable dont les valeurs sont des catégories discrètes dans la case **Independent List**. Par exemple, ces variables pourraient être l'âge et le sexe, ou encore l'âge au premier mariage et le sexe. La syntaxe obtenue est la suivante :

```
MEANS
 TABLES=age BY marital
 /CELLS MEAN COUNT STDDEV .
```

Si on exécute cette syntaxe, on obtient le tableau suivant.

| Report            |       |      |                |
|-------------------|-------|------|----------------|
| Age of Respondent |       |      |                |
| Marital Status    | Mean  | N    | Std. Deviation |
| 1 married         | 46,49 | 794  | 15,158         |
| 2 widowed         | 71,96 | 163  | 10,785         |
| 3 divorced        | 46,19 | 213  | 12,575         |
| 4 separated       | 40,78 | 40   | 11,720         |
| 5 never married   | 31,57 | 285  | 12,288         |
| Total             | 46,23 | 1495 | 17,418         |

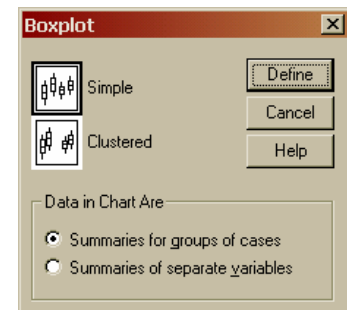
On voit ici que l'âge moyen des personnes mariées est de 46 ans environ, alors que celui des personnes veuves est de près de 72 ans. Celles qui n'ont jamais été mariées ont en moyenne presque 32 ans.

Ceci peut être illustré par le graphique suivant : Cliquez **Graphs→Boxplots**. Vous obtenez la boîte de dialogue suivante :

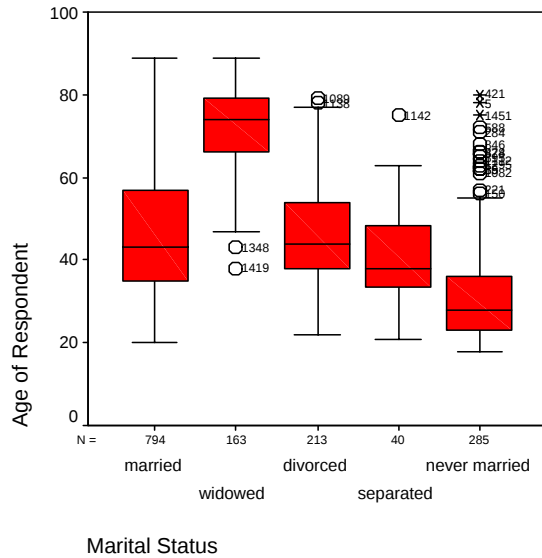
Cliquez sur **Simple** et sur **Summaries of groups of cases**, puis sur **Define**.

Dans la boîte de dialogue qui apparaît, mettez la variable **age** dans la boîte **Variable**, et la variable **Marital Status** dans la boîte **Category Axis**, puis cliquez **Paste**. Vous devriez obtenir la syntaxe suivante.

```
EXAMINE
 VARIABLES=age BY marital
 /PLOT=BOXPLOT/STATISTICS=NONE/NOTOTAL
 /MISSING=REPORT.
```



En l'exécutant, vous obtenez le diagramme de la page suivante qui illustre non seulement les différences de moyennes entre les divers groupes, mais aussi les différences dans la distribution des valeurs.



**Exercice 9.1.** Écrire vos réponses dans un document Word à remettre au plus tard au prochain cours.

1. Examiner les différences d'âge au premier mariage entre les personnes ayant complété des niveaux d'études différents. Tirez les conclusions qui se dégagent en vous souvenant que ces conclusions ne concernent que notre échantillon. Nous n'avons pas encore appris à généraliser ce type de relation à l'ensemble de la population.
2. Examiner les différences de revenu entre hommes et femmes et ajouter à vos conclusions toutes les mises en garde qui s'imposent, compte tenu de la façon dont la variable Revenu est codée.
3. Examiner les différences de revenu entre en fonction des allégeances politiques (**partyid**).
4. Examiner les différences de revenu en fonction des quatre catégories d'âge (**agecat4**)

## Labo 10 La corrélation et la régression

Quand les deux variables sont quantitatives, l'association statistique entre elles prend la forme de la corrélation. Ce terme est synonyme du terme : association statistique entre variables quantitatives.

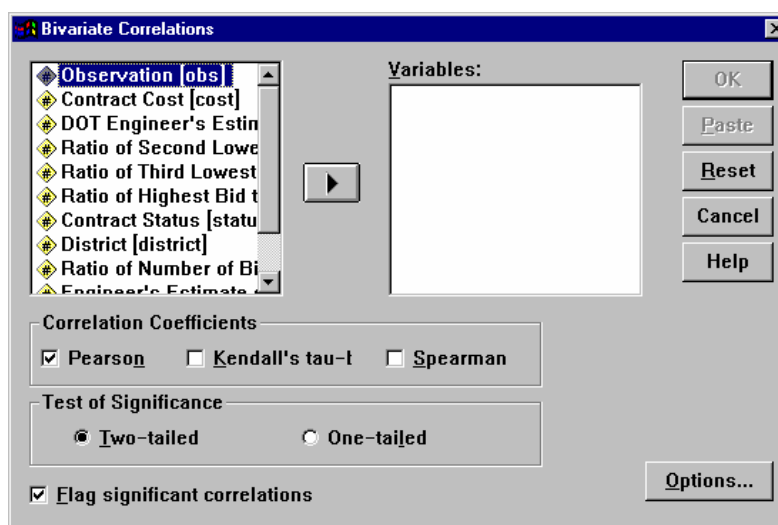
Pour cet exercice, nous utiliserons le fichier intitulé **Road constructions bids** qui est fourni avec SPSS et auquel on accède en cliquant **More Files...** lorsqu'on démarre SPSS. Nous allons étudier les corrélations entre les coûts estimés et les coûts réels de projets de construction de routes entrepris par une agence municipale de transport, désignée dans le fichier par le terme **DOT**, soit **Department of Transport**.

Nous allons effectuer deux opérations distinctes. L'une consiste à produire les coefficients de corrélation entre des variables, et l'autre à dessiner le nuage de points et à obtenir la ligne de régression. SPSS peut produire les coefficients de corrélation entre plusieurs variables prises deux à deux, d'un seul coup. On général on fait cette opération dans un premier temps pour explorer la situation, afin de déceler les relations significatives, puis on analyse avec plus de détails ces relations.

### Calcul des coefficients de corrélation

Pour obtenir les coefficients de corrélation, nous allons exécuter les étapes suivantes.

1. Ouvrez le fichier **Road construction bids**. Pour le trouver, sélectionner **More files...** quand vous ouvrez SPSS. Vous obtenez une liste de fichiers, et celui-ci est dans la liste.
2. Prenez le temps d'examiner les variables présentes dans le fichier, et surtout leur échelle de mesure. Les variables traitent des coûts de certains projets de construction, des coûts estimés, et du nombre de jours de travail nécessaires pour leur exécution.
3. Nous allons examiner la relation entre le coût estimé d'un projet et son coût réel. L'estimé est donné par la variable **dotest**, dont l'étiquette est '**DOT Engineer's Estimate of Construction Cost**', et le coût réel est donné par la variable **cost**, dont l'étiquette est '**Construction cost**'. Dans un premier temps, nous voulons savoir dans quelle mesure les estimés des ingénieurs étaient proches des coûts réels.
4. Sélectionnez : **Analyze → Correlate → Bivariate...**  
Vous obtenez la boîte de dialogue suivante :



5. Placez les variables **Construction cost** et **DOT Engineer's Estimate of Construction Cost** dans l'espace prévu à cet effet à droite.
6. Cliquez **OK** (vous pouvez aussi utiliser la syntaxe si vous préférez). Vous obtenez le tableau suivant :

## Correlations

|                                              |                     | Contract Cost | DOT Engineer's Estimate of Construction Cost |
|----------------------------------------------|---------------------|---------------|----------------------------------------------|
| Contract Cost                                | Pearson Correlation | 1,000         | ,987                                         |
|                                              | Sig. (2-tailed)     | ,000          | ,000                                         |
|                                              | N                   | 235           | 235                                          |
| DOT Engineer's Estimate of Construction Cost | Pearson Correlation | ,987          | 1,000                                        |
|                                              | Sig. (2-tailed)     | ,000          | ,000                                         |
|                                              | N                   | 235           | 235                                          |

\*\* Correlation is significant at the 0.01 level (2-tailed).

Le coefficient de corrélation qui nous intéresse est de 0.987, ce qui est une forte corrélation. Ceci signifie qu'en général, les coûts estimés sont pas mal proche de la réalité : ce sont de bons estimés des coûts réels. Mais ils ne sont pas identiques aux coûts réels pour autant.

Vous aurez sans doute remarqué qu'en plus de donner le coefficient de corrélation (appelé coefficient de Pearson), le tableau vous donne aussi un niveau de signification, et le nombre de cas qui ont été inclus dans le calcul. Le niveau de signification nous dit quel risque de se tromper on prend si on prétend que la relation observée est valable pour l'ensemble de la population étudiée en supposant évidemment que les données que l'on a constituent un échantillon représentatif). Le nombre de cas utilisé est important car il se peut qu'il y ait des données manquantes. Dans notre cas, les 235 données cas du fichiers ont été inclus. Il n'y a pas de données manquantes.

Remarquez aussi qu'il y a une certaine redondance dans le tableau. La corrélation d'une variable avec elle-même est toujours 1. De plus, la corrélation entre x et y est la même qu'entre y et x. Donc, une partie du tableau aurait pu être omise, et certaines versions de SPSS omettent effectivement certaines des cellules redondantes. Ainsi, le tableau suivant contient exactement les mêmes informations que le précédent, rien de moins, car on sait comment remplir toutes les cellules vides.

## Correlations

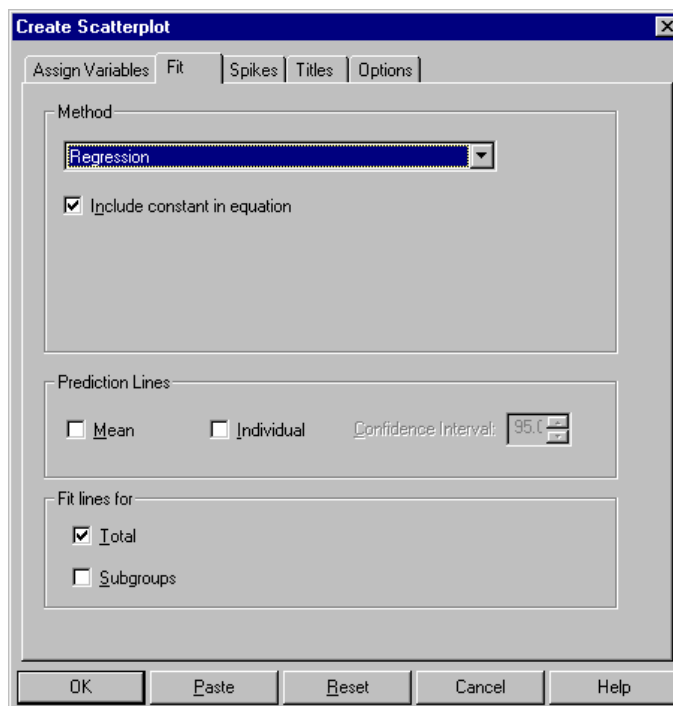
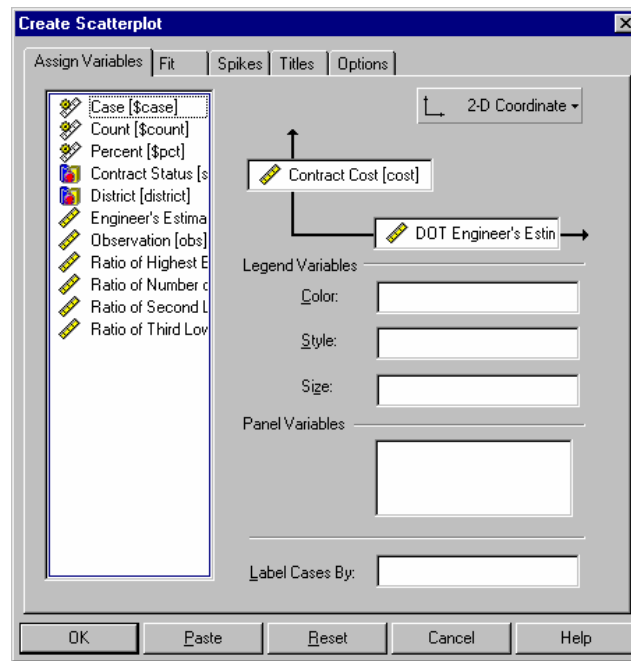
|                                              |                     | Contract Cost | DOT Engineer's Estimate of Construction Cost |
|----------------------------------------------|---------------------|---------------|----------------------------------------------|
| Contract Cost                                | Pearson Correlation |               | ,987                                         |
|                                              | Sig. (2-tailed)     |               | ,000                                         |
|                                              | N                   |               | 235                                          |
| DOT Engineer's Estimate of Construction Cost | Pearson Correlation |               |                                              |
|                                              | Sig. (2-tailed)     |               |                                              |
|                                              | N                   |               |                                              |

\*\* Correlation is significant at the 0.01 level (2-tailed).

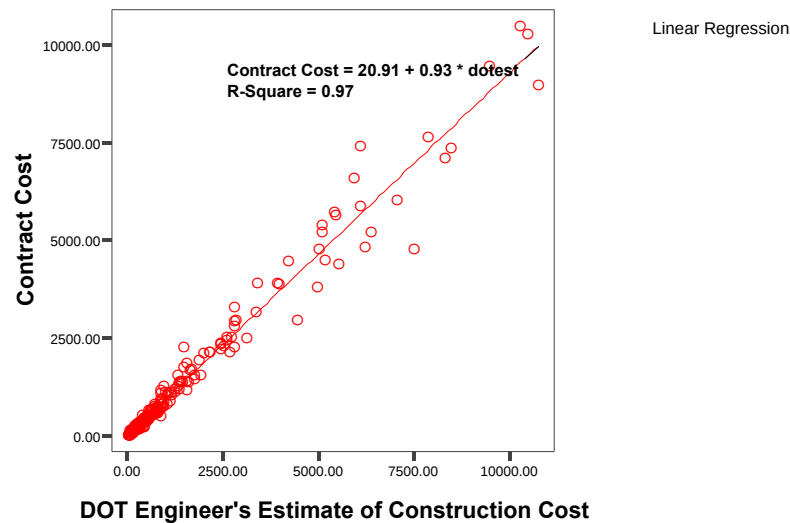
## Illustration graphique et ligne de régression

Les étapes suivantes vont nous permettre d'illustrer la situation.

1. Cliquez **Graphs → Interactive → Scatterplot**.
2. Dans la boîte de dialogue qui en résulte, faites glisser la variable **dotest** vers l'axe horizontal du graphique, and the variable **cost** vers l'axe vertical. La boîte dialogue devrait avoir l'air de celle qui est illustrée à la page suivante
3. Cliquez sur l'onglet dénommé **Fit** dans la partie supérieure de la boîte de dialogue. Vous obtenez une nouvelle boîte de dialogue : assurez-vous que l'option **Regression** a bien été choisie, et que la petite boîte correspondant au mot **Means** n'a pas été sélectionnée. Ceci est illustré à la page suivante.



10. L'onglet **Options** vous permet de choisir plusieurs styles de diagrammes pour le nuage de points. Nous avons choisi 'Classic'.
11. Cliquez **OK**. Vous devriez obtenir le diagramme de la page suivante.



Le diagramme illustre la relation entre les deux variables, et il donne l'équation mathématique de la droite qui exprime la tendance générale. Nous pouvons tirer les conclusions suivantes du graphique :

1. Il y a une forte corrélation entre l'estimé que font les ingénieurs de l'agence de transport, et les coûts réels des projets. La corrélation est de 0.987, donc dans le tableau produit plus haut. Le coût estimé est donc un bon prédicteur des coûts réels d'un projet.
2. Cependant, les ingénieurs de l'agence de transport ont tendance à surestimer légèrement les coûts. L'équation de la régression comporte en effet un coefficient b de 0,93 (moins que 1) et un ajustement de près de 20 \$ (le coefficient a). Cette équation apparaît au haut du diagramme illustrant le nuage de points.
3. Nous constatons aussi que les estimés sont plus précis pour les petits contrats que pour les gros.
4. Pour un projet donné, nous pouvons estimer le coût réel du projet de deux façons : graphiquement d'abord, en trouvant la valeur y qui correspond à la valeur x proposée par les ingénieurs : c'est celle que la ligne nous donne. Ou encore en utilisant l'équation. Ainsi, un projet estimé à 5000 \$ par les ingénieurs coûtera en réalité autour de :

$$20.91 + 0.93 (5000) = 20.91 + 4650 = 4671 \$$$

(Vous aurez sans doute remarqué qu'il s'agit de notre propre estimé des coûts réels, celui que l'on fait à partir de l'estimé des ingénieurs !! Le coût réel exact est donné par les données elles-mêmes, et graphiquement, par le point qui représente un contrat.)

## Exercise 10.1

Ouvrez le fichier **World95**, and examinez les corrélations entre les variables suivantes. Les variables sont désignées par l'étiquette (Value Label) anglaise qu'elles ont dans le fichier. Après avoir produit le tableau des corrélations pour toutes les variables, sélectionnez deux corrélations fortes et une moyenne et faites-en l'analyse. Essayer d'écrire des analyses similaires à celles qui ont été faites ci-haut.

- Average female life expectancy (espérance de vie – femmes)
- People who read (%) (pourcentage d'alphabétisation dans la population)
- Female who read (%) (pourcentage d'alphabétisation des femmes)
- Infant mortality (deaths per 1000 live births) (mortalité infantile)
- Daily calorie intake (calories consommées en moyenne par jour, par personne)
- Birth rate per 1000 people (taux de natalité)



## Labo 11: Estimation et intervalles de confiance

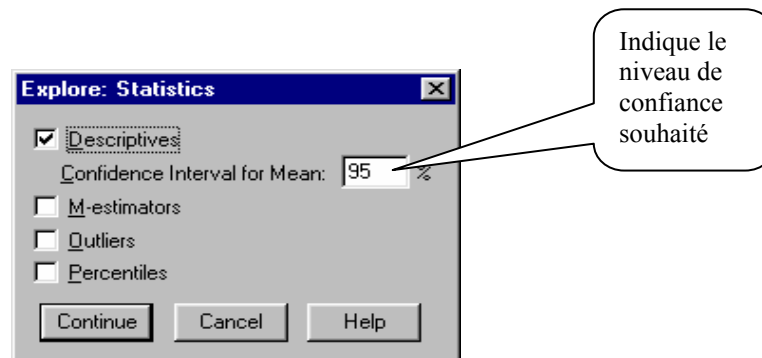
Le but de ce labo est d'apprendre comment produire une procédure d'estimation et en interpréter les résultats. Rappelez-vous que lorsqu'on estime la valeur d'un paramètre à partir d'une statistique, on n'obtient jamais une valeur unique, mais un ensemble de valeurs probables à un certain niveau de confiance. On peut formuler l'ensemble de ces valeurs probables comme un intervalle (dit intervalle de confiance), ou encore comme une valeur ponctuelle accompagnée (toujours) d'une marge d'erreur. La largeur de l'intervalle ainsi que les marges d'erreur dépendent du risque de se tromper que l'on est prêt à prendre, ou encore, inversement, du niveau de confiance qu'on souhaite incorporer dans nos résultats.

### Intervalle de confiance pour la moyenne

SPSS peut calculer les intervalles de confiance pour la moyenne d'une variable quantitative. Ceci est fait par la commande **Explore**, vue dans le laboratoire 3. Nous illustreront la méthode par un exemple tiré du fichier **Employee Data**. Voici comment procéder.

1. Ouvrir le fichier **Employee Data**.
2. Cliquer sur **Analyze**, puis **Descriptive Statistics** puis **Explore...**
3. Dans la boîte que vous obtenez, vous pouvez choisir les variables que vous voulez analyser. Choisissez une variable quantitative, disons la variable **Months Since Hire**, surnommée **jobtime**, et placez-la dans la boîte des variables dépendantes.
4. Vous avez également un bouton appelé Statistics qui vous permet de déterminer ce que vous voulez voir calculer. Cliquez dessus. Vous obtenez la boîte de dialogue montrée dans la figure 11.1. Le mot **Descriptives** est sélectionné, et le niveau de confiance proposé est 95 %. Vous pouvez le changer en 99 % ou en 90 % si vous préférez.

Figure 11.1



5. Cliquez sur **Continue**, puis **Paste** dans la boîte de dialogue **Explore**.

Dans la fenêtre de la syntaxe, vous obtiendrez les commandes suivantes :

```
EXAMINE
 VARIABLES=jobtime
 /PLOT BOXPLOT STEMLEAF
 /COMPARE GROUP
 /STATISTICS DESCRIPTIVES
 /CINTERVAL 95
 /MISSING LISTWISE
 /NOTOTAL.
```

Il y a une commande principale EXAMINE, et plusieurs commandes secondaires. Les deux seules qui soient importantes ici sont la commande VARIABLES qui vous permet de spécifier la variable à analyser, et celle qui détermine l'intervalle de confiance, soit CINTERVAL 95. De sorte que vous pouvez effacer toutes les lignes sauf les suivantes :

```
EXAMINE
 VARIABLES=jobtime
 /CINTERVAL 95.
```

(N'oubliez pas de mettre un point à la dernière ligne).

Si vous faites rouler la commande précédente, vous obtiendrez le même résultat que la commande complète obtenue lorsque vous avez cliqué **Paste** plutôt que **OK**.

Maintenant faite rouler cette commande. Vous obtenez le tableau 11.1.

**Tableau 11.1. Descriptives**

|                   |                                  | Statistic   | Std. Error |
|-------------------|----------------------------------|-------------|------------|
| Months since Hire | Mean                             | 81,11       | ,462       |
|                   | 95% Confidence Interval for Mean | Lower Bound |            |
|                   |                                  | Upper Bound |            |
|                   |                                  | 82,02       |            |
|                   | 5% Trimmed Mean                  | 81,12       |            |
|                   | Median                           | 81,00       |            |
|                   | Variance                         | 101,223     |            |
|                   | Std. Deviation                   | 10,061      |            |
|                   | Minimum                          | 63          |            |
|                   | Maximum                          | 98          |            |
|                   | Range                            | 35          |            |
|                   | Interquartile Range              | 18,00       |            |
|                   | Skewness                         | -,053       | ,112       |
|                   | Kurtosis                         | -1,153      |            |

Examinez la signification des trois premières lignes de ce tableau :

- La moyenne pour cet échantillon est de 81,11 mois. Ceci signifie qu'en moyenne, les employéEs ont été embauchés depuis environs 81 mois (on peut évidemment convertir ce nombre en années et en mois, ou encore en jours).
- L'intervalle de confiance est donné dans les deux lignes ombragées : de 80,20 à 82,02 mois. Ce que cela signifie, c'est que si les individus dans cet échantillon étaient un groupe représentatif d'une plus grande population d'employés, vous estimeriez le temps moyen de travail depuis l'embauche pour cette plus grande population quelque part entre 80,20 mois et 82,02 mois. Mais attention : ceci ne signifie pas que le temps depuis l'embauche des employés se situe quelque part entre 80 et 82 mois, mais bien que *la moyenne* du temps depuis l'embauche, pour toute la population, se situe dans ces limites approximatives.
- L'erreur type est l'écart type divisé par la racine carrée de n, et cette quantité est employée dans le calcul de la marge de l'erreur et elle est donnée à la droite du tableau. Vous n'avez pas besoin de

l'employer, puisque l'intervalle de confiance a été calculé par le programme de SPSS. Mais faites quand même la vérification : la formule de l'intervalle de confiance est donnée par

[moyenne de l'échantillon – 1.96\*l'erreur type ; moyenne de l'échantillon + 1.96\*l'erreur type

Appliquez cette formule. Vous devriez obtenir [80,20 ; 82,02]

- Les mesures restantes ont été vues dans le chapitre sur les statistiques descriptives.

Ces notions employées ci-haut ont été expliquées dans le chapitre sur l'estimation et vous devriez passer en revue ce chapitre afin d'interpréter correctement les résultats donnés par SPSS.

### Remarque sur la taille de l'échantillon

Vous avez sans doute remarqué que l'intervalle de confiance est très petit, c'est-à-dire que la précision de l'estimation est très grande. D'une part il faut noter que notre unité de mesure est le mois. Si le temps était mesuré en jours, l'intervalle de confiance aurait l'air d'être plus grand numériquement (60, plutôt 2 ...) mais ne l'est pas en réalité puisque le premier nombre compte des jours et le deuxième des mois. Cette grande précision est due au fait que l'échantillon est assez grand. Mais si vous recommenciez l'exercice en sélectionnant d'abord un échantillon de 60 ou de 100 personnes, vous verrez que l'intervalle de confiance serait plus grand, c'est-à-dire que le résultat serait moins précis. Rappelez-vous que l'erreur type est donnée par l'écart type divisé par la racine de n. Plus n est grand, plus l'erreur type est petite (et par conséquent l'intervalle de confiance l'est aussi). Mais comme il y a une racine carrée au dénominateur, cette relation inverse n'est pas proportionnelle : un échantillon 4 fois plus grand donne un intervalle 2 fois plus petit et non pas 4 fois plus petit.

### Exercices avec SPSS

1. Écrivez la syntaxe simplifiée vous-même, mais changez le niveau de confiance de 95 % à 99 %. Qu'arrive-t-il à l'intervalle de confiance ? Et si le niveau de confiance était de 90 % ?
2. Écrivez l'énoncé d'estimation comme une phrase complète pour les situations suivantes, supposant que le fichier de données est un groupe représentatif d'une certaine population. Les énoncés que vous écrivez devraient avoir la même forme que ceux des exercices sur l'estimation faits précédemment.
  - a) Estimez le paramètre de la variable **Jobtime** dans le fichier de données **Employee Data**, pour un niveau de confiance de 95 %.
  - b) Estimez-le pour un niveau de confiance de 99 %.
  - c) Estimez-le pour un niveau de confiance de 90 %.
3. Écrivez des énoncés semblables pour la variable **Salary** dans le fichier de données intitulé **University of Florida**.

## Représentation graphique des intervalles de confiance pour la moyenne

Les intervalles de confiance peuvent être représentés graphiquement comme suit.

1. Sous **Graphs**, choisir **Error Bar...**
2. Dans la boîte de dialogue que vous avez, choisir **Simple** et **Summaries of Separate variables**, puis cliquez sur **Define**.
3. Choisissez maintenant une variable quantitative, et placez-la dans l'espace réservé aux variables. Dans la boîte de dialogue, vous verrez un espace où il est indiqué **Bars represent** :. Assurez vous que le choix offert est bien **Confidence interval for the mean**. Vous verrez un espace où indiquer le niveau de confiance désiré : 95 %, ou 90 % ou 99 %.
4. Cliquez **OK** (ou collez la syntaxe et faites-la rouler si vous préférez). Vous allez obtenir un graphique représentant l'intervalle de confiance comme illustré dans la figure 11.2

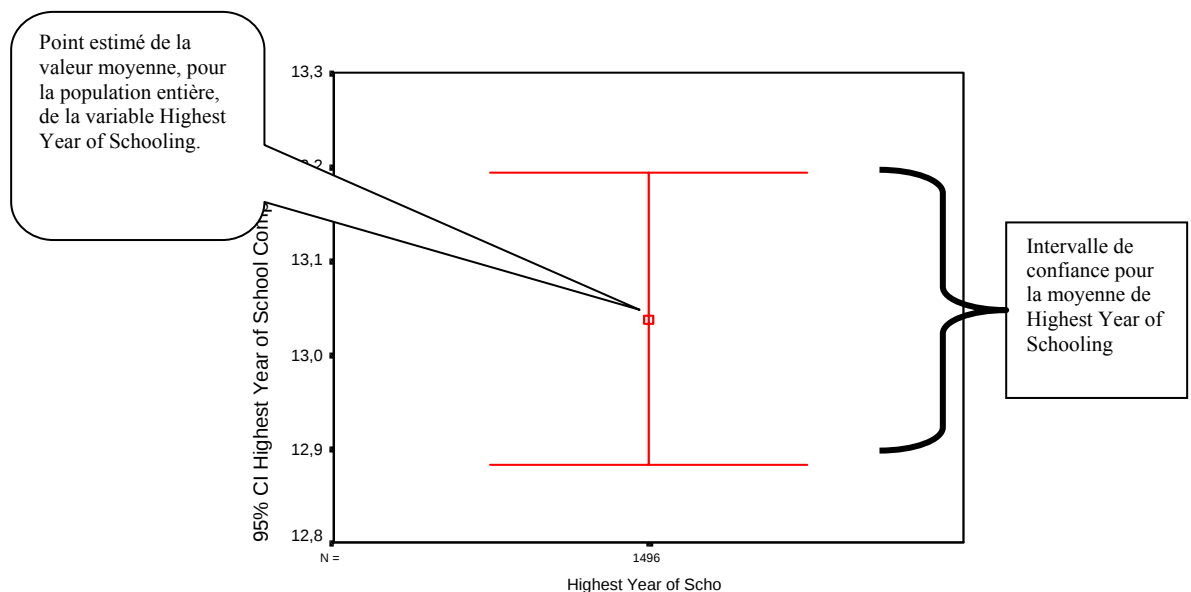


FIGURE 11.2

## Intervalle de confiance pour un pourcentage ou une proportion

Considérons 2 variables :

**cappun** Favor or Oppose Capital Punishment, et  
**attsprts** Attended Sports Event in Last Year  
 qui ont été choisies dans le fichier de données **GSS93 subset**.

Ces deux variables sont des variables qualitatives, et elles sont mesurées à un niveau nominal. Par conséquent, nous ne pouvons pas calculer la moyenne, mais nous pouvons calculer la proportion (ou le

pourcentage) des personnes dans l'échantillon qui sont favorables à la peine de mort (**capital punishment** en anglais), ou qui ont assisté l'année dernière à un événement sportif. Supposons que l'échantillon dans ce fichier de données soit un échantillon aléatoire, nous pouvons alors estimer les pourcentages correspondants dans la population américaine. Mais s'il y a un pourcentage élevé des données manquantes, la fiabilité de telles évaluations est incertaine.

SPSS ne vous fournira pas l'intervalle de confiance du pourcentage, mais donnera le pourcentage calculé dans l'échantillon, en employant la commande **Frequencies**. Vous pouvez déterminer la marge d'erreur au niveau de confiance souhaité en employant la formule donnée dans le chapitre sur l'estimation, formule qui a été incluse dans le fichier Excel intitulé **Calcul des marges d'erreur**. Le tableau suivant donne la valeur approximative des marges d'erreur pour différentes tailles d'échantillon et diverses valeurs du pourcentage calculé dans l'échantillon, à un niveau de confiance de 95% . Comme elles sont approximatives, ces marges d'erreur sont un peu gonflées et elle reflètent la marge d'erreur maximum obtenue pour chaque éventail de pourcentages et de tailles d'échantillon :

| Taille de l'échantillon |     |     |     |     |     |      |      |
|-------------------------|-----|-----|-----|-----|-----|------|------|
| Pourcentage             | 100 | 200 | 400 | 500 | 800 | 1000 | 1500 |
| Autour de 10            | 7   | 5   | 4   | 3   | 3   | 3    | 2    |
| Autour de 20            | 9   | 6   | 5   | 4   | 3   | 3    | 3    |
| Autour de 30            | 10  | 7   | 5   | 5   | 4   | 3    | 3    |
| Autour de 40            | 10  | 7   | 5   | 5   | 4   | 4    | 3    |
| Autour de 50            | 10  | 7   | 5   | 5   | 4   | 4    | 3    |
| Autour de 60            | 10  | 7   | 5   | 5   | 4   | 4    | 3    |
| Autour de 70            | 10  | 7   | 5   | 5   | 4   | 3    | 3    |
| Autour de 80            | 9   | 6   | 5   | 4   | 3   | 3    | 3    |
| Autour de 90            | 7   | 5   | 4   | 3   | 3   | 3    | 2    |

#### Marges d'erreur pour l'estimation d'un pourcentage à un niveau de confiance de 95%.

**Exemple.** Obtenez les fréquences pour la variable **Favor or Oppose Death Penalty for Murder**. Regardez les pourcentages valides. Vous constatez que 77.4 % des réponses valides sont en faveur de la peine de mort en cas de meurtre. Supposant que cet échantillon est représentatif, vous voulez estimer le pourcentage de personnes, dans la population en général, qui sont susceptibles d'être en faveur de la peine de mort en cas de meurtre. Dans ce cas-ci le pourcentage calculé dans l'échantillon est de près de 80 %, et le nombre de réponses valides est 1388, donc très proche de 1500. Le tableau nous donne une marge d'erreur de 3 %. Ainsi, l'énoncé d'estimation devient :

*Sur la base de cet échantillon, nous pouvons estimer que le pourcentage des Américains qui sont en faveur de la peine de mort en cas de meurtre se situe quelque part entre 74.4 % et 80.4 %, pour un niveau de confiance de 95%.*

Ou encore

*Sur la base de cet échantillon, nous pouvons estimer que le pourcentage des Américains qui sont en faveur de la peine de mort en cas de meurtre est de 74.4 %, avec une marge d'erreur de 3 %, au niveau de confiance de 95%..*

## Exercice

4. Considérons 2 variables :

|                 |                                  |    |
|-----------------|----------------------------------|----|
| <b>letdie1</b>  | Allow Incurable Patients to Die, | et |
| <b>scitest4</b> | Humans Evolved From Animals      |    |

qui ont été choisies dans le fichier de données **GSS93 subset**.

Supposant que l'échantillon dans ce fichier de données a été choisi au hasard, essayez de faire une estimation du pourcentage des adultes dans la société américaine qui croient que l'euthanasie devrait être permise pour les patients qui souffrent d'une maladie incurable. Faites également une estimation de ceux qui croient que la théorie de l'évolution (qui affirme que les humains sont le résultat d'une évolution graduelle, à partir de formes de vie moins évoluées) est probablement ou certainement vraie (mettez les catégories ensemble en ajoutant leurs pourcentages). En tenant compte du pourcentage de données manquantes, écrivez un commentaire sur la fiabilité de cette estimation.

## Labo 12: Les tests T de validation d'une hypothèse

Le but de ce laboratoire est d'apprendre comment exécuter une validation d'hypothèse sur la valeur de la moyenne d'une population lorsqu'on a un échantillon, et comment en interpréter les résultats. Cette procédure est appelée **One-Sample T Test** dans le logiciel SPSS. Nous verrons aussi une procédure apparentée, celle qui valide la différence entre les moyennes de deux populations lorsqu'on a deux échantillons indépendants, appelée **Independent-Samples T Test**.

Ces procédures doivent leurs noms à la distribution t, une distribution qui ressemble à la courbe normale, mais qui est plus appropriée quand l'échantillon est petit (moins de 30 individus). Ces procédures sont largement répandues en psychologie, où des expériences sont souvent entreprises sur de petits échantillons. Mais elles sont valides sur de grands échantillons aussi, car à mesure que la taille de l'échantillon croît, la distribution t se rapproche d'une distribution normale.

### Le test d'hypothèse sur la valeur de la moyenne d'une population

**Examen de la méthode.** Dans ce test d'hypothèse, vous voulez valider l'hypothèse que la moyenne  $\mu$  de la population entière diffère d'une certaine valeur, qui est déterminée par une expérience précédente ou par analogie avec une situation semblable. Par exemple, si vous savez par expérience que la note moyenne dans un cours donné est de 77 sur 100 dans un groupe d'écoles, et que vous voulez examiner si une classe spécifique diffère de manière significative de cette moyenne, vous posez :

$$H_0 : \mu = 77$$

$$H_1 : \mu \neq 77$$

Si la moyenne de votre échantillon diffère légèrement de 77, vous n'avez pas une assez bonne raison de rejeter l'hypothèse nulle, car une petite différence entre la moyenne de la population et celle d'un échantillon est explicable par le hasard : un échantillon aléatoire est susceptible en effet de différer légèrement de la population entière. Mais si la différence est grande, vous concluez qu'elle n'est probablement pas due au hasard : si cet échantillon est représentatif, il va refléter le fait que la moyenne de la population est **probablement** différente de 77. Mais comment juger de l'importance de la différence observée? À quelle distance est le point de coupure (appelé *valeur critique*, rappelez-vous) à partir duquel nous pouvons dire : la différence entre la moyenne observée dans l'échantillon et la moyenne supposée de la population est trop grande pour être due au hasard ? Puisque le procédé entier est basé sur la vraisemblance de la conclusion, la réponse dépendra du risque que nous sommes disposés à prendre en tirant nos conclusions. Disons que nous sommes disposés à prendre un risque de 5% de nous tromper. SPSS calculera la probabilité de tomber sur un échantillon tel que celui que vous avez, si l'hypothèse nulle était vraie. Cette probabilité est appelée « niveau de signification ». Si le niveau de signification est plus petit que le risque de 5% que vous avez fixé, vous concluez que vous pouvez prendre ce risque, et vous rejetez l'hypothèse nulle et acceptez l'hypothèse alternative. Si le niveau de signification est plus grand que 5%, vous concluez que le risque est trop grand, et vous concluez que vous n'avez pas d'assez bonnes raisons de rejeter l'hypothèse nulle. Voyons comment exécuter cette procédure concrètement.

### Exemple 1

Supposez que vous voulez évaluer l'hypothèse que l'âge moyen de la population américaine est de 45 ans, et vérifier votre hypothèse en employant l'échantillon aléatoire indiqué dans le fichier de données **GSS93 subset**. Vous avez placé vos hypothèses comme

$$H_0 : \mu = 45$$

et  $H_1 : \mu \neq 45$

Vous lancez la procédure en choisissant

**Analyze → Compare Means → One-Sample T Test...**

Vous obtenez la boîte de dialogue illustrée dans la figure. 12.1. Vous pouvez voir dans la figure que nous avons déjà placé la variable **Age of respondent** dans la boîte **Test Variable(s)** et la valeur que nous voulons valider, appelée **Test Value** dans SPSS, été placée à 45.

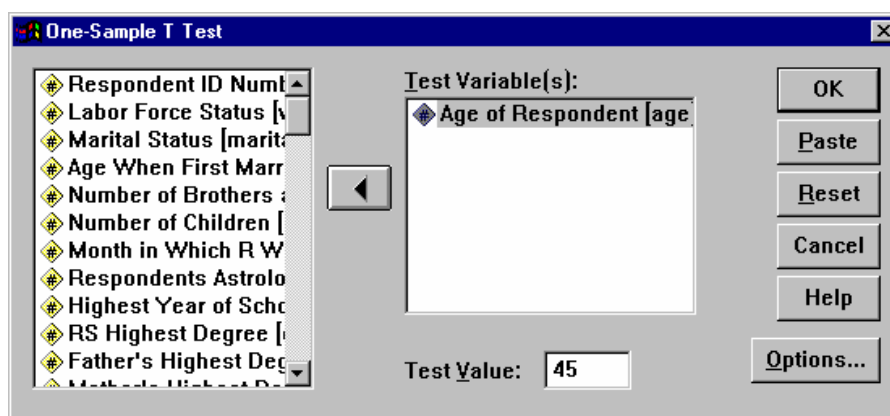


Figure 12.1

Si vous cliquez **OK**, vous obtenez un premier tableau qui vous apprend que la moyenne d'âge pour cet échantillon est de 46,23 ans, et un deuxième tableau que est reproduit ci-bas (tableau 12.1).

**Tableau 12.1.**  
**One-Sample Test**

|                   | T     | df   | Sig.<br>(2-tailed) | Test Value = 45 |                                              |               |
|-------------------|-------|------|--------------------|-----------------|----------------------------------------------|---------------|
|                   |       |      |                    | Mean Difference | 95% Confidence Interval<br>of the Difference |               |
| Age of Respondent | 2,723 | 1494 | <b>,007</b>        | 1,23            | Lower<br>,34                                 | Upper<br>2,11 |

L'information cruciale dans ce tableau est la colonne intitulée le **Sig (2-tailed)**, qui représente le niveau de signification. Vous l'interprétez comme suit :

*Si la moyenne d'âge de votre population est en effet de 45 ans, la probabilité de sélectionner aléatoirement un groupe de 1495 individus dont la moyenne d'âge est de 46.23 ans est de 0.007, ou 0.7%*

Ceci signifie que si la moyenne d'âge de la population générale était effectivement de 45 ans, il serait très peu probable de tomber sur un si gros échantillon dont la moyenne d'âge est de 46,23 ans. En d'autres termes, une différence de 1,23 ans sur un échantillon de cette taille est trop grande pour être due au hasard. Ceci se produirait moins de 1 % des fois. C'est tellement rare qu'il est plus sûr pour vous de conclure plutôt que votre hypothèse est probablement erronée. En faisant un tel raisonnement, vous courez un risque de 0.7 % d'avoir tort, puisque c'est la probabilité d'obtenir un tel échantillon quand la moyenne de la population est de 45 ans. Vous concluez donc que l'hypothèse  $\mu = 45$  doit être remise en question, à la lumière de la moyenne calculée sur cet échantillon. L'hypothèse nulle est ainsi rejetée, avec une probabilité de .007 de se tromper, ce qui constituerait une erreur de Type I.

**Conclusion :  $H_0$  est rejeté puisque le niveau de signification est moins de 0.05. Nous concluons que l'âge moyen de l'ensemble de la population n'est probablement pas de 45 ans.**

Le tableau donne aussi l'intervalle de confiance de la différence entre la moyenne calculée sur l'échantillon et la moyenne supposée, au niveau de confiance de 95%. L'interprétation de cet intervalle est plus compliquée à formuler. Il s'agit dans notre exemple de l'intervalle [0,34 ; 2,11], dont le centre est 1,23. Ceci signifie que 95



fois sur cent, un tel échantillon provient d'une population dont la moyenne pourrait différer d'une distance qui varie entre 0,34 unités et 2,11 unités. Puisque la valeur 0 ne se trouve pas dans cette intervalle, ceci signifie que la possibilité que la moyenne de la population soit égale à 45 est exclue (avec 5 % de chances de se tromper, évidemment).

## Utilisation de la syntaxe

Si au lieu de cliquer **OK** dans l'exemple précédent, vous cliquez **Paste** pour coller la commande dans la fenêtre de la syntaxe, vous obtenez ce qui suit.

```
T-TEST
 /TESTVAL=45
 /MISSING=ANALYSIS
 /VARIABLES=age
 /CRITERIA=CIN (.95) .
```

Vous aurez sans doute remarqué que, comme toutes les commandes, celle-ci est composée d'une commande principale (T-TEST), et de sous-commandes. Les sous-commandes sont séparées par une barre oblique, ne se terminent pas par des points, et sont en retrait vers la droite. Le point clôt l'ensemble de la commande et de ses sous-commandes. Les sous-commandes sont les suivantes :

- TESTVAL=45                qui vous permet de déterminer la valeur à tester.
- VARIABLES=age            qui vous permet de déterminer les variables que vous voulez analyser. Vous pouvez en mettre plusieurs, séparées par des espaces. Ici, on n'a que la variable **age**.
- CRITERIA=CIN (.95)      qui vous permet de déterminer l'intervalle de confiance souhaité (CIN), qui est dans ce cas de 0,95 (remarquez que pour la syntaxe, SPSS utilise des points et non pas des virgules pour le séparateur de décimales).

(Nous ne nous occuperons pas de la commande MISSING pour le moment).

En faisant rouler cette commande, nous obtenons les mêmes résultats que ceux obtenus plus haut. L'avantage de la syntaxe, est que si vous voulez exécuter la procédure à nouveau mais au niveau de confiance de 99 %, il suffit de changer le .95 pour un .99 dans la sous-commande CRITERIA et de la faire rouler à nouveau. L'avantage se fait sentir lorsqu'on exécute de nombreuses commandes, ou qu'on les applique successivement à des bases de données différentes.

## Validation d'une hypothèse portant sur la différence entre deux moyennes

**Examen de la méthode.** Le **Independent-Samples T Test** nous aide à déterminer si deux échantillons, choisis indépendamment, sont susceptibles de provenir de la même population. En d'autres termes, nous supposons que:

L'échantillon 1 provient d'une population avec une moyenne  $\mu_1$ , et

L'échantillon 2 provient d'une population avec une moyenne  $\mu_2$ .

Nous faisons alors l'hypothèse que  $\mu_1 = \mu_2$ , ou, d'une manière équivalente, que

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_1 : \mu_1 - \mu_2 \neq 0$$

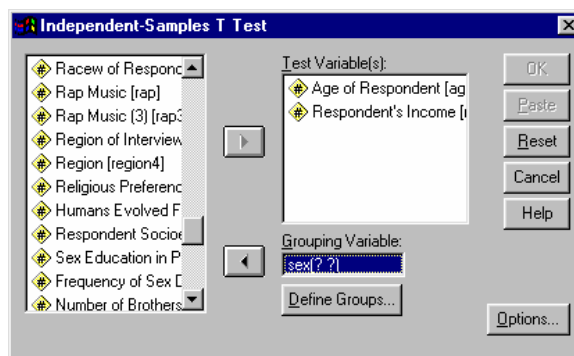
L'hypothèse nulle signifie que les deux échantillons proviennent de populations ayant des moyennes identiques. SPSS calculera la différence des moyennes entre les deux échantillons, et calculera un niveau de signification. L'exemple suivant nous aidera à interpréter les résultats.

## Exemple 2

Nous voulons valider l'hypothèse que la différence entre les hommes et les femmes dans notre échantillon sur les variables : **age** et **rincome91** sont significatifs, c'est-à-dire qu'elles reflètent une vraie différence au niveau de la population entière.

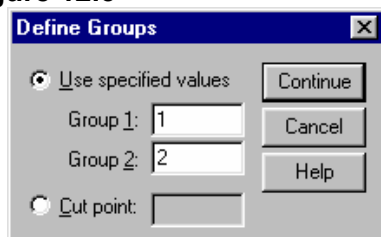
1. Choisissez **Independent-Samples T Test** (dans le menu **Analyse**, puis **Compare Means**). Vous obtenez la boîte de dialogue montrée dans la figure. 12.2.

Figure 12.2



2. Placez les variables **age** et **rincome91** dans les boîtes appropriées comme représenté sur la figure 12.2. Vous devriez réaliser que vous exécutez deux tests d'hypothèse différents en même temps, un pour chacune des variables. SPSS vous permet de faire cela.
3. Placez la variable **sex** dans l'espace étiqueté **Grouping Variable:**. Deux points d'interrogation apparaissent alors, nous permettant de déterminer les deux catégories de cette variable que nous souhaitons comparer. Pour la variable **sex**, il n'y en a que deux. Mais si on voulait comparer les personnes divorcées et les personnes veuves, par exemple, on pourrait le faire en indiquant les codes de leur catégorie, tel que montré au paragraphe suivant.
4. Cliquez sur la boîte **Define Groups...** box. Vous obtenez la boîte de dialogue montrée dans la Figure 12.3.

Figure 12.3



Dans ces boîtes de dialogue, inscrivez 1 pour le premier groupe, celui des hommes, et 2 pour le groupe 2, celui des femmes. Si vous vouliez comparer les personnes veuves et les personnes divorcées, on aurait inscrit **marital** au lieu de **sex**, et dans les groupes on aurait choisi les groupes 2 (veufs) et 3 (divorcés).

5. Cliquez sur **Continue**, puis sur **OK**. Vous obtenez le tableau 12.2 (certaines des colonnes du tableau dont nous n'avons pas besoin tout de suite ont été supprimées).

**Tableau 12.2.**  
**Independent Samples T Test**

|                     |                             | t-test for Equality of Means |                 |                 |                                           |       |
|---------------------|-----------------------------|------------------------------|-----------------|-----------------|-------------------------------------------|-------|
|                     |                             | t                            | Sig. (2-tailed) | Mean Difference | 95% Confidence Interval of the Difference |       |
|                     |                             |                              |                 |                 | Lower                                     | Upper |
| Age of Respondent   | Equal variances assumed     | -1,697                       | <b>,090</b>     | -1,54           | -3,33                                     | ,24   |
|                     | Equal variances not assumed | -1,708                       | <b>,088</b>     | -1,54           | -3,32                                     | ,23   |
| Respondent's Income | Equal variances assumed     | 7,470                        | <b>,000</b>     | 2,59            | 1,91                                      | 3,28  |
|                     | Equal variances not assumed | 7,488                        | <b>,000</b>     | 2,59            | 1,91                                      | 3,27  |

L'information cruciale ici est le niveau de signification calculé, dénoté par **Sig. (2-tailed)**. Les résultats sont calculés dans deux cas : le cas où les variances des sous-populations des hommes et des femmes sont identiques, et le cas où les variances ne sont pas identiques. SPSS offre des tests pour déterminer si ces variances sont identiques ou pas, mais la discussion de ces tests ne sera pas abordée pour le moment. Comme règle pratique, considérez que les variances sont égales : les erreurs de Type I seront alors moins probables.

Concrètement, voici comment interpréter les résultats de ce tableau.

La variable **Age of Respondent**: Dans ce cas-ci, l'hypothèse nulle est qu'il n'y a aucune différence entre les âges des hommes et des femmes dans l'ensemble de la population. Nous supposons que la population des hommes et des femmes ont la même variance pour la variable âge. Si nous affirmons que la différence entre les hommes et les femmes est significative, nous prenons un risque de 9% d'avoir tort. C'est parce que la différence entre leurs moyennes est très petite : 1.54 ans. Il est trop risqué de dire qu'une si petite différence pour cet échantillon indique une vraie différence au niveau de la population entière. Nous devrions plutôt expliquer la différence par le hasard : il est plus probable que les échantillons choisis indépendamment montrent une telle différence, même si ils viennent de la même population. *Par conséquent, dans ce test,  $H_0$  est acceptée. Nous concluons que nous n'avons pas une raison suffisante de rejeter l'hypothèse que les hommes et les femmes dans la population entière ont le même âge moyen.*

La variable **Respondent's Income** : Ceci constitue un test d'hypothèse différent du précédent, puisque la variable est différente. L'hypothèse nulle est qu'il n'y a aucune différence entre les revenus des hommes et ceux des femmes dans l'ensemble de la population. Les revenus sont regroupés en 22 catégories, codées 1 à 22. La différence moyenne entre les scores moyens des hommes et des femmes est 2.59 (les scores réfèrent aux catégories, et non pas au montant du revenu en dollars). C'est une différence relativement importante : le revenu moyen des hommes se situe en moyenne deux catégories au-dessus de celui des femmes. Les résultats de SPSS confirment cette interprétation : nous prenons un risque qui est pratiquement nul (arrondi à moins de 0.000) quand nous affirmons que cette différence est significative. Par conséquent, nous pouvons conclure qu'il y a une vraie différence entre les revenus des hommes et des femmes au niveau de la population entière, pas simplement pour ce groupe de 1500 personnes. *Par conséquent, dans ce test,  $H_0$  est rejetée et  $H_1$  est acceptée avec un risque de se tromper plus petit que 0.0005 (car si la quatrième décimale était 5 ou plus, on aurait arrondi la probabilité à 0,001).*

### **Règle pratique pour interpréter le niveau de signification calculé**

Si le niveau de signification calculé par SPSS, **Sig.(2-tailed)**, est inférieur au niveau de signification que nous avons fixé (c'est-à-dire le risque que nous sommes disposés à prendre), nous rejetons  $H_0$  et acceptons  $H_1$ . Si **Sig.(2-tailed)** est plus grand que le niveau de signification que nous avons fixé nous acceptons  $H_0$  et rejetons  $H_1$ .

## **Utilisation de la syntaxe**

Comme pour les autres procédures, on peut utiliser la syntaxe. Dans l'exemple précédent, nous devons utiliser la syntaxe suivante.

```
T-TEST
 GROUPS=sex(1 2)
 /MISSING=ANALYSIS
 /VARIABLES=age rincom91
 /CRITERIA=CIN(.95) .
```

La commande est la même que pour le test impliquant un seul échantillon, mais la première sous-commande, **GROUPS**, nous indique que l'on compare deux groupes, et elle détermine ces groupes. Si on voulait comparer les veufs et les divorcés (hommes ou femmes, indistinctement), on aurait :

```
T-TEST
 GROUPS=marital(2 3)
 /MISSING=ANALYSIS
 /VARIABLES=age rincom91
 /CRITERIA=CIN(.95) .
```

Les autres sous-commandes sont les mêmes que dans l'exemple précédent. Vous avez toujours le choix de dactylographier ces commandes directement en observant les règles de la syntaxe, plutôt que de travailler avec les menus.

## **II. Exercices additionnels**

Utilisez le fichier de données **GSS93 subset**. Pour chaque exercice, écrivez vos conclusions en phrases complètes. Spécifiez  $H_0$  et  $H_1$  chaque fois que nécessaire, en mots et en équations, pour que l'interprétation des conclusions soit limpide.

**(Note importante :** La partie « formulation » de ces exercices pourrait vous sembler oiseuse, mais elle est très efficace pour bien intérioriser les connaissances et les conceptualiser clairement).

1. Validez l'hypothèse selon laquelle les hommes et les femmes ont tendance à se marier à des âges différents.
2. Validez l'hypothèse selon laquelle ceux ou celles qui ont un diplôme universitaire ont tendance à se marier plus tard que ceux ou celles qui n'en ont pas.
3. Sélectionnez un échantillon aléatoire de 100 personnes et refaites les mêmes deux tests précédents. Qu'est-ce qui arrive aux intervalles de confiance ? Aux erreurs types (**standard error** qui, rappelez-vous, sont les écarts types des distributions d'échantillonnage correspondantes) ? Est-ce que les conclusions diffèrent du cas où vous utilisez toutes les données ? (Notez que vous n'obtiendrez pas tous les mêmes résultats car vous n'aurez sans doute pas sélectionné les mêmes échantillons....)
4. Si on conclut que les variables mesurant l'éducation universitaire (**degree2**) et le sexe (**sex**) ont toutes les deux un impact sur la variable âge au premier mariage, peut-on dire laquelle a le plus grand impact ? Pour cela, exécutez la commande **Compare means**, en choisissant le niveau d'éducation comme

premier niveau (**Layer 1**) et la variable sexe comme deuxième niveau (**Layer 2**), tel que montré en classe. Ou encore, exécutez la syntaxe suivante :

```
MEANS
 TABLES=agewed BY degree2 BY sex
 /CELLS MEAN COUNT STDDEV .
```

et interprétez le tableau qui en résulte. Concluez votre analyse avec des énoncés de la forme suivante :

« La variable X a pour effet de retarder l'âge moyen du mariage de ..... années, alors que la variable Y a pour effet de retarder l'âge moyen du mariage de ..... années. L'effet combiné des deux variables a pour effet de retarder l'âge moyen du mariage de .... ..années : en effet, les ..... qui ont un diplôme universitaire se marient en moyenne, ..... années plus tard que les ..... qui n'en ont pas ».

## Labo 13 : Le test du Chi-deux

**Rappel** : Le test du Chi-deux est un test de validation d'hypothèses. Il s'applique aux tableaux croisés. On calcule une statistique qui mesure l'écart entre une situation théorique où il n'y aurait pas d'association statistique, et une la situation observée. Cette statistique suit une distribution connue, qui dépend du nombre de catégories des variables étudiées.

On pose donc :

$H_0$  : Il n'y a pas d'association statistique entre les variables.

$H_1$  : Il y en a.

S'il n'y avait aucune association statistique dans la population d'où provient l'échantillon, il y aurait quand même des différences observées au niveau de l'échantillon, mais seulement dans 5 % des cas ces différences produiraient une statistique qui dépasserait un certain seuil. Dans 1 % des cas, la statistique dépasserait un autre seuil plus élevé. Sur cette base, on peut effectuer les calculs qui vont nous amener à accepter l'hypothèse nulle ou à la rejeter. SPSS va en effet nous donner la probabilité qu'on obtienne la valeur du Chi-deux observée sur l'échantillon s'il n'y avait aucune association statistique au niveau de la population. On rejette l'hypothèse nulle si la probabilité est plus petite que le seuil qu'on s'est fixé.

### Exemple

En utilisant le fichier **GSS93 subset**, nous allons valider l'hypothèse que les femmes et les hommes ont des attitudes différentes concernant la peine de mort. Nous supposons évidemment que l'échantillon est représentatif. La généralisation que nous voulons faire n'est valide qu'à cette condition.

Nous posons donc :

$H_0$  : Les hommes et les femmes appuient la peine de mort dans les mêmes proportions

$H_1$  : Les hommes et les femmes appuient la peine de mort dans des proportions différentes.

Nous retiendrons un seuil de signification de 5%.

Ouvrons le fichier SPSS **GSS93 subset** et effectuons la procédure **Crosstabs** apprise au Labo 8. Placez la variable Respondent's Sex dans l'espace réservé pour les lignes du tableau, et la variable **Favor or Oppose Death Penalty for Murder (cappun)** dans l'espace réservés pour les colonnes. Cliquez sur le bouton **Statistics** et cochez la case correspondante au Chi-deux (**Chi-squared**). Demandez aussi les pourcentages par ligne, et vous les obtiendrez. La syntaxe obtenue est la suivante.

```
CROSSTABS
 /TABLES=sex BY cappun
 /FORMAT= AVALUE TABLES
 /STATISTIC=CHISQ
 /CELLS= COUNT ROW .
```

Vous aurez remarqué que nous avons mis la sous-commande du Chi-deux en caractères gras pour attirer votre attention sur la façon de l'écrire. Quand vous exécutez cette commande, vous obtenez évidemment le tableau croisé que vous avez vu précédemment, mais vous obtenez aussi le tableau du Chi-deux. Nous reproduisons les deux tableaux.

**Respondent's Sex \* Favor or Oppose Death Penalty for Murder Crosstabulation**

|                  |          |                           | Favor or Oppose<br>Death Penalty for<br>Murder |          |        |
|------------------|----------|---------------------------|------------------------------------------------|----------|--------|
|                  |          |                           | 1 Favor                                        | 2 Oppose | Total  |
| Respondent's Sex | 1 Male   | Count                     | 502                                            | 105      | 607    |
|                  |          | % within Respondent's Sex | 82,7%                                          | 17,3%    | 100,0% |
|                  | 2 Female | Count                     | 572                                            | 209      | 781    |
|                  |          | % within Respondent's Sex | 73,2%                                          | 26,8%    | 100,0% |
| Total            |          | Count                     | 1074                                           | 314      | 1388   |
|                  |          | % within Respondent's Sex | 77,4%                                          | 22,6%    | 100,0% |

Vous constatez que la différence entre les hommes et les femmes est près de 10 points de pourcentage (9,5 % de différence plus exactement entre le pourcentage de femmes et d'hommes qui appuient la peine capitale dans le cas d'un meurtre). Cette différence semble grande, mais est-elle assez grande pour dire qu'il y a une différence au niveau de toute la population, pas seulement l'échantillon ? Le tableau suivant nous donne la réponse.

**Chi-Square Tests**

|                              | Value     | df | Asymp. Sig.<br>(2-sided) | Exact Sig.<br>(2-sided) | Exact Sig.<br>(1-sided) |
|------------------------------|-----------|----|--------------------------|-------------------------|-------------------------|
| Pearson Chi-Square           | 17,470(b) | 1  | ,000                     | ,000                    | ,000                    |
| Continuity Correction(a)     | 16,934    | 1  | ,000                     |                         |                         |
| Likelihood Ratio             | 17,800    | 1  | ,000                     |                         |                         |
| Fisher's Exact Test          |           |    |                          |                         |                         |
| Linear-by-Linear Association | 17,458    | 1  | ,000                     |                         |                         |
| N of Valid Cases             | 1388      |    |                          |                         |                         |

a Computed only for a 2x2 table

b 0 cells (,0%) have expected count less than 5. The minimum expected count is 137,32.

## Lecture du tableau du Chi-deux

**Le niveau de signification.** Une seule cellule du tableau nous importe pour le moment, c'est celle que nous avons indiquée par une trame de fond. Elle donne la valeur du niveau de signification relatif au Chi-deux de Pearson, qui est dans ce cas 0,000. Rappelez-vous que cela n'est pas égal à zéro : ce nombre signifie en effet que les trois premières décimales sont zéro. Il pourrait y avoir des décimales qui ne sont pas égales à zéro plus loin. Comme on arrondi vers le haut si la quatrième décimale est 5 ou plus, on peut en conclure que **le niveau de signification est plus petit que 0,0005**.

Ceci signifie que : Si les hommes et les femmes étaient pour la peine de mort dans les mêmes proportions dans la population dans son ensemble, il y aurait une probabilité plus petite que 0,0005 (i.e. moins de 5 chances sur 10 000) qu'on obtienne un échantillon comme celui que l'on a, avec des différences de 9,5 % entre les deux.

**Conclusion :** Nous acceptons  $H_1$  avec un risque d'erreur presque nul, et rejettons  $H_0$  comme étant très peu probable (moins de 0,0005).

**La valeur du Chi-deux.** Le tableau nous donne une valeur de 17,470. En soi cela ne nous dit probablement pas grand-chose. Cette valeur résulte du calcul effectué avec la formule du Chi-deux (essayez de l'obtenir vous-même à l'aide de la feuille de calcul Excel fournie durant le cours). Pour un échantillon de cette taille, une valeur égale à 17,470 ou plus grande qu'elle ne se retrouverait que moins de 5 fois sur 10 000 si les hommes et les femmes de la population avaient les mêmes attitudes par rapport à la peine de mort.

**Le degré de liberté.** Il est donné dans la colonne df (pour degrees of freedom). Il est calculé par la formule :

$$\text{Degré de liberté} = (n-1)*(m-1)$$

Où n et m sont les nombres de catégories dans les deux variables étudiées. Dans notre cas,  $n=2$  et  $m=2$ , donc le degré de liberté  $= (2-1)*(2-1) = 1*1 = 1$ . Le degré de liberté a un sens technique : il nous dit dans laquelle des distributions il faut regarder pour déterminer la probabilité d'obtenir un tel échantillon. Nous n'avons pas besoin d'utiliser ce nombre directement puisque SPSS fait les calculs pour nous, mais il est bon de comprendre que plus il y a de degrés de liberté, plus la valeur du Chi-deux des échantillons varie de la valeur théorique.

**Les mises en garde.** Vous remarquerez que SPSS a calculé qu'aucune des cellules n'a une fréquence théorique plus petite que 5. En effet, les calculs de probabilités relatifs au Chi-deux ne sont valables que si cette condition est remplie : les fréquences théoriques de chacune de cellule doit être égale à 5 ou plus.

**Note sur la taille de l'échantillon.** Il n'est pas étonnant d'avoir une probabilité aussi petite compte tenu de la taille de l'échantillon. En effet, plus un échantillon aléatoire est grand, plus il donnera des pourcentages qui se rapprochent de ceux de la population. Un petit échantillon pourrait s'en éloigner bien plus. Pour le vérifier, sélectionnez successivement un échantillon aléatoire de 30 personnes, puis de 50 personnes, puis de 100 puis de 200 personnes, et faites le test du Chi-deux. Comparer les diverses valeurs du Chi-deux que vous obtenez ainsi que les divers niveaux de signification. Vous verrez comment la taille de l'échantillon affecte ces diverses statistiques.

## Exercice

En utilisant le même fichier, **GSS93 subset**, Déterminer si la différence entre les pourcentages d'hommes et de femmes qui se prévalent de leur droit de vote est significative. Refaire l'exercice avec un échantillon aléatoire de 100 personnes. Recommencer l'exercice avec les niveaux d'éducation plutôt que le sexe.



## UTILISER EXCEL POUR FAIRE DES CALCULS ÉLÉMENTAIRES

### 1. Les formules dans Excel

Le logiciel Excel, ainsi que les autres tableurs similaires, permettent d'inscrire dans les cellules des formules qui sont calculées automatiquement. Ces formules peuvent contenir des nombres, ainsi des *références relatives* ou des *références absolues* à d'autres cellules (termes expliqués plus bas). Toutes les formules doivent commencer par le signe = . Ensuite, on peut dactylographier la formule, ou cliquer sur les cellules qu'on veut inclure (ceci sera montré en classe). En voici quelques exemples :

|                 |                                                                        |
|-----------------|------------------------------------------------------------------------|
| = (5*6) + 2     | Multiplie 5 par six d'abord, puis additionne 2.                        |
| = 5*(6+2)       | Additionne 6 +2, puis multiplie par 5.                                 |
| = A1*A2         | Multiplie le contenu de la cellule A1 par le contenu de la cellule A2. |
| = \$A\$1*\$A\$2 | Multiplie le contenu de la cellule A1 par le contenu de la cellule A2. |

Vous aurez remarqué que les deux dernières formules sont écrites différemment, mais que le résultat du calcul est le même. Qu'en est-il exactement ?

La troisième formule comporte une référence *relative* aux cellules A1 et A2, alors que la quatrième formule comporte une référence *absolue* aux mêmes cellules. La différence entre les deux types de référence paraît lorsqu'on copie les formules, tel qu'expliqué dans ce qui suit.

### 2. Copier des formules à référence relative

Quand on sélectionne une cellule et qu'on tape simultanément **Ctrl** et **C**, la formule est copiée dans la mémoire vive de l'ordinateur. Si on déplace le curseur dans une autre cellule et qu'on tape **Ctrl** **V**, la formule est alors recopiée dans cette dernière cellule. Mais attention : si les références dans la formule sont relatives, la formule est alors modifiée de la façon illustrée dans l'exemple suivant :

**Exemple.** Supposons qu'on ait les données du tableau ci-contre. La cellule D7 contient la formule = A1+A2, et les références sont relatives.

Si on recopie cette formule dans la cellule E7 qui est juste à sa droite, les A deviendront des B. Si on la recopie deux cases plus loin, les A deviendront des C. Si on la recopie trois lignes plus bas, les numéros de cellules qui apparaissent dans la formule seront majorés de trois unités.

En d'autres termes, les cellules utilisées dans le calcul subissent le même déplacement que la cellule où on inscrit le résultat.

The screenshot shows an Excel spreadsheet titled 'Microsoft Excel - Classeur1'. The data is organized as follows:

|    | A | B  | C | D       | E | F | G |
|----|---|----|---|---------|---|---|---|
| 1  | 1 | 10 |   |         |   |   |   |
| 2  | 3 | 23 |   |         |   |   |   |
| 3  | 2 | 34 |   |         |   |   |   |
| 4  | 6 | 54 |   |         |   |   |   |
| 5  | 4 | 67 |   |         |   |   |   |
| 6  | 8 | 12 |   |         |   |   |   |
| 7  | 7 | 43 |   | = A1+A2 |   |   |   |
| 8  | 9 | 26 |   |         |   |   |   |
| 9  | 1 | 27 |   |         |   |   |   |
| 10 |   |    |   |         |   |   |   |
| 11 |   |    |   |         |   |   |   |

The formula bar at the top shows the formula in the selected cell: = A1+A2. The status bar at the bottom indicates 'Feuil1', 'Feuil2', 'Feuil3', and 'NUM'.

#### Exercice 1.1

Créez le tableau illustré ci-dessus, dactylographiez la formule illustrée à la case D7, et presser ENTER. (Note : c'est important de taper Enter, sans quoi, tout clic de la souris va modifier la formule que vous avez écrite...). Maintenant cliquez sur D7, copiez la cellule, et collez la successivement dans les cellules indiquées et examinez le résultat obtenu.

Copiez la dans E7; résultat : formule obtenue : \_\_\_\_\_ nombre obtenu : \_\_\_\_\_

Copiez la dans F7; résultat : formule obtenue : \_\_\_\_\_ nombre obtenu : \_\_\_\_\_  
 Copiez la dans G7; résultat : formule obtenue : \_\_\_\_\_ nombre obtenu : \_\_\_\_\_  
 Copiez la dans E12; résultat : formule obtenue : \_\_\_\_\_ nombre obtenu : \_\_\_\_\_

**À présent, veuillez sauvegarder votre document sous le titre : votre\_nom\_de\_famille\_Ex1 .**

### 3. Copier des formules à référence absolue.

#### Exercice 1.2

Recommencer le même exercice en utilisant les références absolues (un signe de \$ avant chaque lettre et avant chaque chiffre). Comment les formules sont-elles modifiées ?

Quel est le résultat du calcul dans chacun des 4 cas de l'exercice précédent ?

### 4. Les commandes 'Recopier à droite' et 'Recopier vers le bas'

Ces commandes sont utilisées pour recopier une formule dans un ensemble de cellules qui se suivent verticalement ou horizontalement. Il faut d'abord écrire une formule dans une cellule, puis sélectionner cette cellule ainsi que les cellules qui la suivent verticalement. En cliquant *Recopier vers le bas* dans le menu Edition, la formule est alors recopiée dans toutes les cellules sélectionnées. L'effet est similaire lorsqu'on recopie à droite. Si les références dans la formule sont relatives, elles seront ajustées automatiquement, alors que les références absolues ne seront pas modifiées. Ces commandes vont grandement faciliter le calcul des mesures descriptives, tel qu'illustré ci-bas.

#### Exercice 1.3

Les données suivantes représentent les notes obtenues dans un classe. Recopiez-les dans la colonne B d'une feuille Excel, en commençant par la troisième ligne. Dans la première ligne de la colonne B, écrivez simplement Note obtenue, et dans la deuxième ligne, écrivez X, tel qu'illustré.

Effectuez les calculs suivant en utilisant des formules :

Dans la cellule B13, calculer la somme des notes obtenues en utilisant la formule =SOMME(B3:B12).

(Remarque : plusieurs façons de produire cette formule seront illustrées en classe. On peut utiliser la commande Insertion, Formule, ou encore dactylographier la formule, ou enfin cliquer sur le signe  $\Sigma$ ).

Dans la cellule B14, inscrivez le nombre de données, soit 10.

Dans la cellule B15, calculer la moyenne des notes par la formule =B13/B14. Inscrivez le mot 'Moyenne' dans la cellule A15.

Inscrivez la formule =B3\*B3 dans la case C3. Recopier la formule vers le bas jusqu'à la ligne 12.

Inscrivez X au carré dans la case C2.

En vous inspirant des étapes précédentes, calculez la somme des carrés des notes par l'entremise d'une formule à la case C13.

Calculez l'écart-type des notes, en tenant compte qu'il s'agit d'un échantillon. La formule est :

|    | A | B            | C | D | E |
|----|---|--------------|---|---|---|
| 1  |   | Note obtenue |   |   |   |
| 2  |   | X            |   |   |   |
| 3  |   | 67           |   |   |   |
| 4  |   | 78           |   |   |   |
| 5  |   | 89           |   |   |   |
| 6  |   | 93           |   |   |   |
| 7  |   | 68           |   |   |   |
| 8  |   | 75           |   |   |   |
| 9  |   | 56           |   |   |   |
| 10 |   | 81           |   |   |   |
| 11 |   | 79           |   |   |   |
| 12 |   | 72           |   |   |   |
| 13 |   |              |   |   |   |
| 14 |   |              |   |   |   |

écart type de l'échantillon :  $s = \sqrt{\frac{\sum (x_i - \bar{X})^2}{n-1}} = \underline{\hspace{2cm}}$

### 5. Calcul d'une moyenne pondérée

Supposez maintenant que vous avez 5 classes A, B, C, D et E, dont la taille et les moyennes des notes obtenues par les étudiants sont données par :

| Classe | N  | Moyenne | Pondération       |
|--------|----|---------|-------------------|
| A      | 26 | 86      | <u>          </u> |
| B      | 20 | 75      | <u>          </u> |
| C      | 30 | 70      | <u>          </u> |
| D      | 12 | 95      | <u>          </u> |
| E      | 28 | 80      | <u>          </u> |

Calculez la moyenne pondérée des notes pour les cinq classes prises ensemble, en calculant d'abord le poids de chaque classe, tel que cela a été illustré par le professeur.

Moyenne pondérée :

## CALCUL DE LA CORRÉLATION ET DE LA DROITE DE RÉGRESSION À L'AIDE D'EXCEL

L'utilisation d'un tableur peut nous aider à mieux comprendre la logique du calcul du coefficient de corrélation de Pearson et de la droite de régression. Nous utiliserons Excel, mais n'importe quel tableur peut être utilisé car nous ne ferons appel qu'à des fonctions relativement simples.

Rappelons que la droite de régression est celle qui passe au centre du nuage de points qui représente les données relatives à deux variables quantitatives, c'est-à-dire par le point  $(\bar{X}, \bar{Y})$ , et qui épouse le mieux la tendance observée dans le nuage de points. Mais comment déterminer la droite qui représente le mieux le nuage de points ? Quel critère utiliser pour dire qu'une droite épouse mieux la tendance qu'une autre ? Pour définir cette droite, nous utiliserons la méthode dite « des moindres carrés » que nous allons expliquer dans ce qui suit.

### La méthode des moindres carrés

Le but de notre démarche est de trouver une ligne droite qui représente le mieux la tendance observée dans le nuage de points. Il faudrait donc que cette droite « colle » aux points du nuage le plus possible, et non seulement qu'elle passe par le centre de ce nuage (qui est le point dont les coordonnées sont les moyennes de X et de Y).

Il y a plusieurs façons de définir une telle droite. Pour identifier la meilleure, il faut se rappeler que cette droite servira aussi à *estimer* la valeur de la variable dépendante correspondant à une valeur donnée de la variable indépendante. En d'autres termes, si on a une certaine valeur  $x_i$ , cette droite devrait estimer la valeur  $y_i$  correspondante, et ce de la meilleure façon possible. Donc, elle devrait minimiser les écarts entre la valeur estimée et les valeurs que l'on retrouve dans nos données. Or si on travaille avec les écarts eux-mêmes, les écarts positifs vont annuler les écarts négatifs et nous ne pourrions pas déterminer la meilleure droite. Nous allons donc recourir au même stratagème mathématique que nous avons utilisé pour calculer l'écart type : nous avons mis les écarts au carré, de façon à avoir à faire avec des quantités qui sont toutes positives et qui ne s'annulent pas mutuellement. On va s'y prendre de la même façon ici : nous allons déterminer la droite qui *minimise la somme des carrés de la distance entre les valeurs observées et les valeurs estimées*. C'est pour cela que cette méthode de calcul de la droite de régression s'appelle la *méthode des moindres carrés*.

Supposons donc que nous ayons des valeurs pour la variable indépendante X :

$$x_1, x_2, x_3, \dots, x_n,$$

et les valeurs correspondantes pour la variable dépendante Y :

$$y_1, y_2, y_3, \dots, y_n.$$

Supposons à présent que la droite recherchée est donnée par  $y = a + bx$ . Quelles sont les valeurs de a et de b qui vont faire que cette droite est la « meilleure », c'est-à-dire qu'elle représente la tendance du nuage de point mieux qu'aucune autre ? Nous avons retenu comme critère pour déterminer cette droite celui des moindres carrés. Si on utilise  $\hat{y}$  pour désigner la valeur de Y estimée par la droite en question, il faudrait donc que la somme des écarts ( $\hat{y} - y$ ) mis au carré soit la plus petite possible. On peut démontrer à l'aide du calcul différentiel que ceci sera réalisé si cette droite par le centre du nuage  $(\bar{X}, \bar{Y})$  donc si

$$a = \bar{Y} - b\bar{X}$$

et si

$$b = \frac{(\sum XY) - n\bar{X}\bar{Y}}{(\sum X^2) - n\bar{X}^2}$$

La formulation mathématique de b ci-haut est celle qui est donnée par beaucoup de manuels. Mais si on veut être plus précis, il faudrait écrire :

$$b = \frac{(\sum x_i y_i) - n \bar{X} \bar{Y}}{(\sum x_i^2) - n \bar{X}^2}$$

(la sommation étant faite sur l'indice i qui prend les valeurs 1, 2, 3, ..., n puisque l'on a n données).

Ces deux équations pour a et b vont nous permettre de calculer nous mêmes la droite de régression.

Or ce calcul peut être effectué assez facilement par un tableur tel que Excel. Il suffira d'écrire toutes les étapes successives du calcul dans des colonnes différentes du tableur, en utilisant la fonction **Recopier vers le bas** tel que montré dans un cours précédent.

Ainsi, si la première colonne contient les valeurs  $x_i$  de la variable X et la deuxième les valeurs  $y_i$  de Y (commençant à la deuxième ligne, pour laisser la première pour les titres des colonnes), on peut créer les colonnes suivantes :

|         |   |   |       |    |
|---------|---|---|-------|----|
| Colonne | 1 | 2 | 3     | 4  |
| Titre   | X | Y | $X^2$ | XY |

Les données suivront dans les lignes suivantes. On utilisera les fonctions pour effectuer le calcul. Par exemple, le calcul de  $X^2$  se fera en inscrivant

$$= A2 * A2$$

dans la colonne C2, puis en recopiant vers le bas. Au bas des données, on additionnera les X et les Y pour obtenir ensuite leurs moyennes, ainsi que les  $X^2$  et les XY. Puis on inscrira la formule pour le b dans une nouvelle cellule, et celle du a dans une autre.

Note : **Les exercices suivants peuvent être effectués chez vous, car ils nécessitent le logiciel Excel seulement. Cependant, les données doivent être préalablement recopiées du fichier SPSS Road Construction Bids qui se trouve sur les ordinateurs du Labo. Pour les recopier, il suffit de sélectionner ces données puis de faire un copier/coller dans la première cellule d'une feuille de calcul Excel : le nombre de cellules requises s'ajustera automatiquement.**

### Exercice 1 :

1. Faites le calcul de l'équation de régression pour les variables : **Construction Cost** (variable dépendante) et **DOT's Engineers Estimate** (variable indépendante), et vérifiez que vous obtenez les mêmes réponses que lorsque SPSS effectue les calculs.
2. Calculez le coefficient de corrélation pour ces deux variables à l'aide d'Excel et de la formule suivante (rappelez vous que l'estimé de Y est donné par la formule  $a + bX$ ) :

$$r^2 = 1 - \frac{\sum (Y - \hat{Y})^2}{\sum (Y - \bar{Y})^2}$$

### Exercice 2 :

Refaites l'exercice précédent avec une autre paire de variables du même fichier ou du fichier **WORLD95**. Vous n'avez pas besoin de saisir les formules, qui peuvent être simplement recopiées.

**DÉPOSEZ votre document Excel** comportant les exercices dans le Fichier Gourou Cours → SOC 4206 → Dépôt.

## ESTIMATION

### Exemple d'un énoncé:

Le sondage effectué sur 1030 individus a montré que 37 % des adultes canadiens tirent leurs informations internationales de la télévision. Ces résultats sont précis à  $\pm 4$  %, et sont fiables 19 fois sur 20. (données fictives)

La population

LA TAILLE DE L'ÉCHANTILLON

LA VARIABLE MESURÉE

LA STATISTIQUE MESURÉE

LE PARAMÈTRE ESTIMÉ

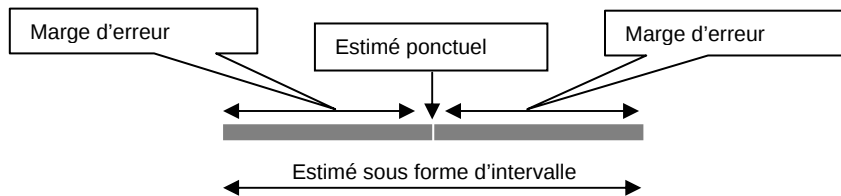
LA MARGE D'ERREUR

NIVEAU DE CONFIANCE

PROBABILITÉ D'ERREUR

## Estimation d'une proportion

|                                                         |                                                                    |
|---------------------------------------------------------|--------------------------------------------------------------------|
| Si on veut un niveau de confiance de<br><br><b>90 %</b> | La marge d'erreur est de<br><br>$\pm 1.64 \sqrt{\frac{p(1-p)}{n}}$ |
| Si on veut un niveau de confiance de<br><br><b>95 %</b> | La marge d'erreur est de<br><br>$\pm 1.96 \sqrt{\frac{p(1-p)}{n}}$ |
| Si on veut un niveau de confiance de<br><br><b>99 %</b> | La marge d'erreur est de<br><br>$\pm 2.58 \sqrt{\frac{p(1-p)}{n}}$ |



**Estimation d'une moyenne**

Le sondage effectué sur **1030 individus** a montré que les **adultes canadiens regardent la télévision en moyenne 4,2 h par jour**. Ces résultats sont précis à **± 6 minutes près**, et sont fiables **19 fois sur 20**. (données fictives)

|                                                         |                                                             |
|---------------------------------------------------------|-------------------------------------------------------------|
| Si on veut un niveau de confiance de<br><br><b>90 %</b> | La marge d'erreur<br><br>$\pm 1.64 \frac{\sigma}{\sqrt{n}}$ |
| Si on veut un niveau de confiance de<br><br><b>95 %</b> | La marge d'erreur<br><br>$\pm 1.96 \frac{\sigma}{\sqrt{n}}$ |
| Si on veut un niveau de confiance de<br><br><b>99 %</b> | La marge d'erreur<br><br>$\pm 2.58 \frac{\sigma}{\sqrt{n}}$ |



### ***Effets de la taille de l'échantillon***

Si  $n = 100$ , la formule comporte 10 au dénominateur.

Pour obtenir une marge d'erreur 2 fois plus petite, il faudrait diviser par 20, soit avoir un échantillon 4 fois plus grand.

Donc :

**Il faut multiplier la taille de l'échantillon par 4  
pour diminuer la marge d'erreur de moitié**

### ***Calcul de la taille d'échantillon nécessaire***

Pour une proportion, si on a déterminé la marge d'erreur  $m$ , on peut isoler la valeur de  $n$  qui va produire cette marge d'erreur. Comme la marge d'erreur est maximale quand  $p = 0.5$ , on obtient :

$$\text{Taille de l'échantillon } n = \left( \frac{1.96 * 0.5}{m} \right)^2$$

Similairement, pour les moyennes, on obtient :

$$\text{Taille de l'échantillon } n = \left( \frac{1.96 * \sigma}{m} \right)^2$$

ESSAYER DE FAIRE CE CALCUL AVEC EXCEL

**Exercices sur l'estimation****NOM :** \_\_\_\_\_**I. Interprétation des énoncés d'estimation**

Veuillez lire les énoncés suivants et en tirer les informations demandées. Les données sont fictives.

1. Une étude effectuée sur un échantillon aléatoire de 430 femmes adultes dans la région métropolitaine de Montréal a montré que 73 % des femmes préfèrent utiliser leur auto pour ce rendre au travail. Les résultats sont précis à  $\pm 4\%$ , 19 fois sur 20.

Variable étudiée : \_\_\_\_\_  
 Population étudiée : \_\_\_\_\_  
 Taille de l'échantillon : \_\_\_\_\_ Statistique mesurée : \_\_\_\_\_  
 Valeur ponctuelle estimée du paramètre : \_\_\_\_\_ Intervalle: \_\_\_\_\_  
 Marge d'erreur : \_\_\_\_\_  
 Probabilité d'erreur : \_\_\_\_\_ Niveau de confiance : \_\_\_\_\_

2. Sur la base d'une enquête faite à l'UQÀM, il a été établi que les étudiantEs prennent en moyenne 43 minutes ( $\pm 11$  minutes) pour se rendre à l'Université. Ces résultats sont fiables 9 fois sur 10.

Variable étudiée : \_\_\_\_\_  
 Population étudiée : \_\_\_\_\_  
 Taille de l'échantillon : \_\_\_\_\_ Statistique mesurée : \_\_\_\_\_  
 Valeur ponctuelle estimée du paramètre : \_\_\_\_\_ Intervalle: \_\_\_\_\_  
 Marge d'erreur : \_\_\_\_\_  
 Probabilité d'erreur : \_\_\_\_\_ Niveau de confiance : \_\_\_\_\_

**Question** : Est-ce que cet énoncé signifie que les étudiants prennent quelque part entre 32 et 54 minutes pour se rendre à l'UQÀM ? \_\_\_\_\_ (Oui/Non). Expliquez votre réponse en 2 lignes.

3. Une enquête auprès d'un échantillon représentatif de volontaires dans les organisations de comté d'un parti politique a montré que les volontaires font en moyenne 7 heures et 32 minutes de bénévolat chaque semaine. Ces résultats sont précis à  $\pm 45$  minutes, avec un risque d'erreur de 10%.

Variable étudiée : \_\_\_\_\_  
 Population étudiée : \_\_\_\_\_  
 Taille de l'échantillon : \_\_\_\_\_ Statistique mesurée : \_\_\_\_\_  
 Valeur ponctuelle estimée du paramètre : \_\_\_\_\_ Intervalle: \_\_\_\_\_  
 Marge d'erreur : \_\_\_\_\_  
 Probabilité d'erreur : \_\_\_\_\_ Niveau de confiance : \_\_\_\_\_

4. Les Québécois préfèrent passer leurs vacances au Québec. Un sondage récent où 2045 personnes ont été interviewées par téléphone a démontré que 69 % d'entre eux préoyaient rester au Québec l'été prochain. La marge d'erreur est de  $\pm 2\%$  avec un niveau de confiance de 95 %.

Variable étudiée : \_\_\_\_\_  
 Population étudiée : \_\_\_\_\_  
 Taille de l'échantillon : \_\_\_\_\_ Statistique mesurée : \_\_\_\_\_  
 Valeur ponctuelle estimée du paramètre : \_\_\_\_\_ Intervalle: \_\_\_\_\_  
 Marge d'erreur : \_\_\_\_\_  
 Probabilité d'erreur : \_\_\_\_\_ Niveau de confiance : \_\_\_\_\_

5. Les étudiantEs dépensent en moyenne entre 4.45 \$ and 5.15 \$ à la cafétéria durant l'heure du dîner. C'est du moins ce qui ressort d'un sondage effectué auprès de 560 étudiants et étudiantes, et les résultats sont fiables 9 fois sur 10.

Variable étudiée : \_\_\_\_\_  
 Population étudiée : \_\_\_\_\_  
 Taille de l'échantillon : \_\_\_\_\_ Statistique mesurée : \_\_\_\_\_  
 Valeur ponctuelle estimée du paramètre : \_\_\_\_\_ Intervalle: \_\_\_\_\_  
 Marge d'erreur : \_\_\_\_\_  
 Probabilité d'erreur : \_\_\_\_\_ Niveau de confiance : \_\_\_\_\_

6. Considérez les deux énoncés suivants, qui se réfèrent à la question 3 ci-haut.

- a) Nous estimons, avec un niveau de confiance de 90 %, que chacun des volontaires du parti passe entre 6 heures et 47 minutes, et 8 heures et 17 minutes chaque semaine à travailler pour le parti.
- b) Nous estimons, avec un niveau de confiance de 90 %, que les volontaires du parti passent en moyenne entre 6 heures et 47 minutes, et 8 heures et 17 minutes chaque semaine à travailler pour le parti.

Quelle est la différence entre ces deux énoncés ? Lequel traduit correctement l'énoncé de la question 3 ?

## **Formulation d'énoncés d'estimation**

Écrivez une phrase complète qui formule un estimé du paramètre pour les exemples suivants. Ceci nécessitera le calcul de la marge d'erreur à l'aide d'Excel. Complétez aussi les espaces laissés blancs.

7. Variable étudiée : Le fait de fumer des cigarettes.  
 Population étudiée : Tous les employés d'une grande compagnie.  
 Taille de l'échantillon : 238 personnes  
 Statistique étudiée : Le pourcentage des fumeurs et fumeuses  
 Statistique mesurée : 29 %  
 Valeur ponctuelle estimée du paramètre : \_\_\_\_\_  
 Intervalle: \_\_\_\_\_  
 Marge d'erreur : \_\_\_\_\_  
 Probabilité d'erreur : \_\_\_\_\_  
 Niveau de confiance : 95 %

Énoncé

---



---



---



---

8. Variable étudiée : Le comportement des conducteurs aux arrêts.  
Population étudiée : Tous les conducteurs de voitures dans une ville.  
Taille de l'échantillon : 1200 personnes  
Statistique étudiée : Le pourcentage de ceux et celles qui font un arrêt complet  
Statistique mesurée : 90 %  
Valeur ponctuelle estimée du paramètre : \_\_\_\_\_  
Intervalle: \_\_\_\_\_  
Marge d'erreur : \_\_\_\_\_  
Probabilité d'erreur : \_\_\_\_\_  
Niveau de confiance : 95 %  
Énoncé : \_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

9. Variable étudiée : Heures de travail rémunéré par semaine  
Population étudiée : Les étudiants de 1ère année du Bac à l'UQÀM.  
Taille de l'échantillon : 900 personnes  
Statistique observée : Nombre d'heures travaillées par semaine  
Statistique mesurée : 15 heures  
Écart type : 3 heures (suggestion : convertir en minutes)  
Valeur ponctuelle estimée du paramètre : \_\_\_\_\_  
Intervalle: \_\_\_\_\_  
Marge d'erreur : \_\_\_\_\_  
Probabilité d'erreur : \_\_\_\_\_  
Niveau de confiance : 95 %

Énoncé : \_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

## Exercices sur la distribution normale

1. Supposons que le poids des nouveaux-nés dans une maternité soit distribué normalement avec une moyenne de 3.5 kg et un écart type de 0.5 kg. Calculer :
  - a) le pourcentage de nouveaux-nés pesant au-dessus de 4 kg;
  - b) Le pourcentage de nouveaux-nés pesant entre 3.5 kg et 4 kg;
  - c) Le pourcentage de nouveaux-nés pesant plus de 5 kg;
  - d) Le pourcentage de nouveaux-nés pesant moins de 2 kg;
  - e) Le pourcentage de nouveaux-nés pesant moins de 2.3 kg
  - f) Le pourcentage de nouveaux-nés pesant plus de 4.6 kg.

2. Les énoncés suivants sont logiquement équivalents. Ce sont des façons diverses de dire la même chose. Tous ces énoncés renvoient à la valeur  $z = 1$

**Pour  $z = 1$**

1. L'aire sous la courbe normale standardisée entre 0 et 1 est de 0.3413 unités.
2. Dans  $N(72,4)$ , le pourcentage des données qui tombent entre les valeurs 72 et 76 est de 34.13 %
3. Si une population est distribuée normalement avec une moyenne de 72 et un écart type de 4 unités, le pourcentage de données entre 72 et 76 est de 34.13 %
4. Dans une population dont la distribution est  $N(72,4)$ , le pourcentage de données plus grand que 76 est de  $(50 - 34,13) = 15.87$
5. Si vous pigez au hasard un individu dans une population qui est distribuée normalement  $N(72,4)$ , il y a environ 16 % de chances que son score soit 76 ou plus.

Pour chacune des valeurs suivantes de  $z$ , écrivez cinq énoncés similaires qui soient équivalents entre eux. Valeurs de  $z$  :  $z = 1.6$  ;  $z = 0.8$  ;  $z = 1.96$  ;  $z = -1.6$ .

**Pour  $z = 1.6$**  (Donc le  $x$  correspondant dans  $N(72, 4)$  est égal à  $72 + (1.6 * 4) = 78,4$ )

1. L'aire sous la courbe normale standardisée entre 0 et 1.6 est de \_\_\_\_\_ unités.
2. Dans  $N(72,4)$ , le pourcentage des données qui tombent entre les valeurs 72 et 78,4 est de \_\_\_\_\_ %
3. Si une population est distribuée normalement avec une moyenne de 72 et un écart type de 4 unités, le pourcentage de données entre 72 et 78,4 est de \_\_\_\_\_ %
4. Dans une population dont la distribution est  $N(72,4)$ , le pourcentage de données plus grand que 78,4 est de  $(50 - \text{_____}) = \text{_____}$
5. Si vous pigez au hasard un individu dans une population qui est distribuée normalement  $N(72,4)$ , il y a environ \_\_\_\_\_ % de chances que son score soit 78,4 ou plus.

**Pour  $z = 0.8$**  (Donc le  $x$  correspondant dans  $N(72, 4)$  est égal à  $72 + (0.8 * 4) = \text{_____}$ )

- 1.
- 2.
- 3.
- 4.
- 5.

**Pour  $z = 1.96$**

**Pour  $z = -1.6$**

## MODÈLE D'EXAMEN FINAL

Vous avez trois heures pour compléter les réponses à cet examen. La **partie I** se fera à la main, sans accès au livre ou au notes. Elle devrait prendre moins d'une demi-heure. Vous commencerez la partie II quand vous aurez remis la partie I. Pour la **partie II**, vous avez le droit d'utiliser le livre, et tout autre matériel d'enseignement, incluant les notes de cours et les labos qui ont été distribués. Vous devez écrire les réponses dans un document Word, et dans ce cas les tableaux ou graphiques de SPSS que vous voulez utiliser doivent être recopiés dans ce document. Il serait préférable d'imprimer votre texte afin d'éviter les conséquences des accidents informatiques ! Mais si vous les mettez dans le dossier Dépôt qui lui-même est dans le dossier du cours SOC 4206, sur le serveur Gourou, je l'imprimerai moi-même sur le champs, et vous demanderai de vérifier que toutes les pages ont bien été imprimées avant que vous partiez.

**IMPORTANT** : Recopiez le fichier de données SPSS qui sera dans le fichier Documents (dans SOC 4206 sur Gourou) sur le Bureau de l'ordinateur, et écrivez les réponses à l'examen dans un document que vous aurez sauvegardé sur le Bureau aussi. Quand vous aurez fini, si vous n'avez pas imprimé vos réponses, placez votre document dans le dossier Dépôt. Nommez votre document ainsi :

*Nom\_de\_familleInitiale\_final*. Exemple : si vous vous appelez Pierre Trudeau votre document sera nommé : TrudeauP\_final.

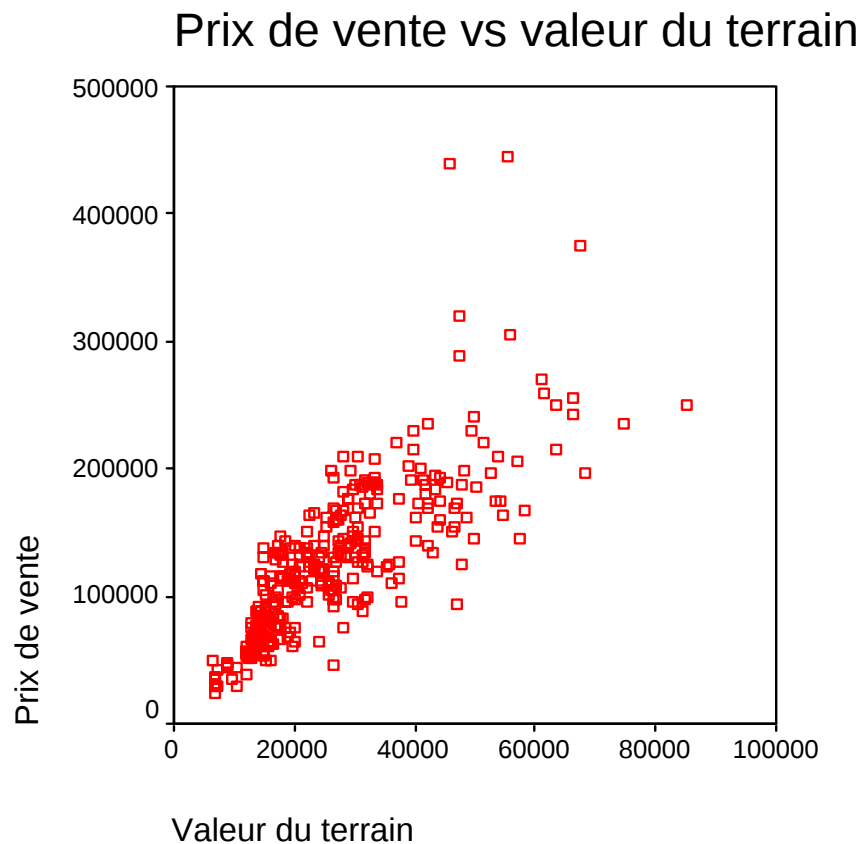
**Suggestions** : 1. Sauvegardez votre document fréquemment !!!

2. Il n'est pas nécessaire de mettre tous les tableaux et graphiques : seulement ceux qui ajoutent quelque chose à la compréhension du texte. Les conclusions doivent être formulées en phrases complètes qui incluent les résultats quantitatifs. Le corrigé de ce modèle d'examen sera disponible la semaine prochaine et nous en discuterons en classe afin que les critères d'évaluation soient clairs pour tout le monde.



**Partie I**

- I. Expliquez en quelques lignes quelles sont les limites et les avantages de la moyenne et de la médiane pour représenter la tendance centrale d'un ensemble de données quantitatives.
- II. Écrivez la formule de l'intervalle de confiance d'un estimé de la moyenne, puis d'une proportion, aux deux niveaux de confiance de 95% et 99%.
- III. Considérez le diagramme de dispersion suivant, qui met en relation la valeur du terrain d'une maison et son prix de vente.



1. Tracez manuellement la ligne de régression.
2. Estimez manuellement le prix de vente moyen d'une maison dont le terrain vaut 40 000 \$.
3. Choisissez une maison qui s'éloigne un peu de la droite de régression et indiquez sur le graphique la différence entre son prix de vente estimé par la droite de régression et son prix de vente réel.

## ***Partie II : Travail sur SPSS***

- I. Analysez les données se rapportant à l'âge au premier mariage, en décrivant l'effet du sexe, du niveau d'éducation, de la variable intitulée 'race', et de la religion pour cet échantillon, chacune de ces variables indépendantes étant prise individuellement.
- II. Créez une variable intitulée : année de naissance, et examinez si il y a une corrélation ou une association statistique entre l'année de naissance et l'âge au premier mariage.
- III. Choisissez une association statistique observée à la question I, et discutez en détail dans quelle mesure elle est vraie pour l'ensemble de la population dont provient cet échantillon, à supposer que ce soit un échantillon aléatoire.
- IV. Analysez les données se rapportant au fait de voter ou pas en 1992, en décrivant l'effet du sexe, du niveau d'éducation, de la variable intitulée 'race', et de la religion pour cet échantillon, chacune de ces variables indépendantes étant prise individuellement.
- V. Choisissez une association statistique observée à la question IV, et discutez en détail dans quelle mesure elle est vraie pour l'ensemble de la population dont provient cet échantillon, à supposer que ce soit un échantillon aléatoire.
- VI. Choisissez un échantillon aléatoire de 100 personnes, et calculez l'âge moyen des individus de cet échantillon ainsi que le pourcentage d'entre eux qui ont voté en 1992. Sur la base de cet échantillon, écrivez deux énoncés pour estimer l'âge moyen de la population, puis le pourcentage de gens qui ont voté aux élections de 1992 au niveau de toute la population.

## RÉPONSES AU MODÈLE D'EXAMEN FINAL

- I. **Question.** Analysez les données se rapportant à l'âge au premier mariage, en décrivant l'effet du sexe, du niveau d'éducation, de la variable intitulée 'race', et de la religion pour cet échantillon, chacune de ces variables indépendantes étant prise individuellement.

**Les données :** Le fichier analysé, intitulé GSS93 subset, est un sous-ensemble des données recueillies lors de l'enquête sociale générale en 1993 aux Etats-Unis. L'échantillon contient 1500 cas, mais il ne semble pas que ce soit un échantillon représentatif car la proportion de femmes et d'hommes diffère grandement de celle de la population générale.

**L'âge au premier mariage.** Les individus de cet échantillon qui se sont mariés l'ont fait pour la première fois à un âge moyen de 22,79 ans, ce qui correspond à 22 ans et 288 jours environ, soit 22 ans et 9 mois et demie environ. L'écart type est de 5 ans. La plus jeune personne à se marier avait 13 ans, et un individu de l'échantillon s'est marié pour la première fois à 58 ans.

**L'effet de la variable sexe.** Les femmes de notre échantillon se marient plus tôt que les hommes. En effet, on peut lire sur le tableau que les femmes se marient à un âge moyen de 21,84 ans et les hommes à un âge moyen de 24,16 ans, l'écart étant de 2 ans et 4 mois environ.

**L'effet de l'obtention d'un diplôme universitaire.** Les personnes qui ont obtenu un diplôme universitaire ont eu tendance à se marier environ 3 ans plus tard, en moyenne, que les non diplômés. En effet la moyenne d'âge au premier mariage pour les premiers est de plus de 25 ans, alors que celle des seconds est d'environ 22 ans.

**L'effet de la variable de classification raciale.** Les blancs et les noirs semblent ne pas trop différer quant à l'âge du premier mariage (22,71 ans vs 22,87 ans respectivement). Les personnes classées 'autres' se marient en moyenne un peu plus tard, tel qu'illustré dans le tableau suivant :

**Tableau 1. Age au premier mariage en fonction de la classification 'raciale'.**

| Race of Respondent | Mean  | N    | Std. Deviation |
|--------------------|-------|------|----------------|
| 1 white            | 22,71 | 1029 | 4,923          |
| 2 black            | 22,87 | 119  | 5,733          |
| 3 other            | 24,28 | 54   | 5,329          |
| Total              | 22,79 | 1202 | 5,033          |

**L'effet de la religion.** La religion semble être un facteur qui affecte l'âge moyen du mariage. Le tableau 2 montre une différence entre les catholiques (23,63 ans) et les protestants (22,25 ans), qui sont les deux groupes religieux les plus nombreux dans cet échantillon. Les autres groupes religieux semblent se marier un peu plus tard, mais leurs effectifs dans cet échantillon sont beaucoup plus réduits.

**Tableau 2. L'âge au premier mariage en fonction de l'appartenance religieuse.**

| Religious Preference | Mean  | N    | Std. Deviation |
|----------------------|-------|------|----------------|
| 1 Protestant         | 22,25 | 787  | 5,014          |
| 2 Catholic           | 23,63 | 265  | 5,004          |
| 3 Jewish             | 25,65 | 23   | 4,141          |
| 4 None               | 23,32 | 95   | 5,015          |
| 5 Other              | 25,42 | 26   | 3,657          |
| Total                | 22,78 | 1196 | 5,032          |

- II. Créez une variable intitulée : année de naissance, et examinez si il y a une corrélation ou une association statistique entre l'année de naissance et l'âge au premier mariage.

La variable « Année de naissance » a été créée en soustrayant l'âge du répondant de l'année où l'enquête a été menée, 1993. La corrélation entre l'année de naissance et l'âge au premier mariage est de  $-0.083$ , soit une corrélation négative très faible. Même si elle est significative (c'est-à-dire qu'elle se généralise à l'ensemble de la population) cette corrélation est très faible et n'a donc pas de valeur explicative : l'année de naissance n'explique que  $(-0.083)^2$ , soit moins de 1 % de la variation de l'âge au premier mariage. On ne peut donc pas conclure que, pour cet échantillon, l'appartenance à des générations plus vieilles explique le mariage à un âge plus jeune.

#### Corrélations

|                        |                        | Age When First Married | année de naissance |
|------------------------|------------------------|------------------------|--------------------|
| Age When First Married | Corrélation de Pearson | 1                      | -.083**            |
|                        | Sig. (bilatérale)      | .                      | .004               |
|                        | N                      | 1202                   | 1199               |
| année de naissance     | Corrélation de Pearson | -.083**                | 1                  |
|                        | Sig. (bilatérale)      | .004                   | .                  |
|                        | N                      | 1199                   | 1495               |

\*\* . La corrélation est significative au niveau 0.01 (bilatéral).

- III. Choisissez une association statistique observée à la question I, et discutez en détail dans quelle mesure elle est vraie pour l'ensemble de la population dont provient cet échantillon, à supposer que ce soit un échantillon aléatoire.

Examinons la relation entre l'âge au premier mariage et le sexe. Pour savoir si la relation observée sur l'échantillon se généralise à toute la population, il faut effectuer un test t. Nous posons :

**L'hypothèse nulle** : Il n'y a aucune différence entre l'âge moyen au premier mariage des hommes et des femmes.

**L'hypothèse alternative** : L'âge moyen au premier mariage des hommes et des femmes est différent.

On obtient le tableau 3, reproduit ci-bas. Nous avons supprimé les colonnes dont nous n'avons pas besoin pour cette analyse.

Tableau 3. Test t pour l'égalité de l'âge au premier mariage des hommes et des femmes

|                             | t     | df     | Sig.<br>(2-tailed) | Mean<br>Difference |
|-----------------------------|-------|--------|--------------------|--------------------|
| Equal variances assumed     | 8,066 | 1200   | ,000               | 2,32               |
| Equal variances not assumed | 8,085 | 1064,6 | ,000               | 2,32               |

Que la variance de l'âge au mariage des femmes et des hommes soit égale ou pas, la conclusion est la même : c'est l'hypothèse alternative qui est acceptée. Ceci signifie qu'on peut affirmer, avec une probabilité presque nulle de se tromper (moins de 0,0005, soit moins de 0,05 %) qu'il y a une différence entre l'âge au premier mariage des hommes et des femmes, en supposant que l'échantillon soit représentatif.

IV. Analysez les données se rapportant au fait de voter ou pas en 1992, en décrivant l'effet du sexe, du niveau d'éducation, de la variable intitulée 'race', et de la religion pour cet échantillon, chacune de ces variables indépendantes étant prise individuellement.

#### Analyse de la participation au vote en 1992

On constate tout d'abord qu'environ 68,8 % des individus de l'échantillon ont déclaré avoir voté en 1992, 28 % ont déclaré ne pas avoir voté, 2,3% ont déclaré ne pas être éligibles, et un tout petit nombre (6 personnes) ont refusé de répondre à cette question. Huit autres cas sont des données manquantes. Ces résultats sont consignés dans le tableau 4.

Tableau 4. Participation au vote en 1992

|                | Frequency | Percent | Valid<br>Percent | Cumulative<br>Percent |
|----------------|-----------|---------|------------------|-----------------------|
| Valid 1 voted  | 1032      | 68,8    | 69,2             | 69,2                  |
| 2 did not vote | 420       | 28,0    | 28,2             | 97,3                  |
| 3 not eligible | 34        | 2,3     | 2,3              | 99,6                  |
| 4 refused      | 6         | ,4      | ,4               | 100,0                 |
| Total          | 1492      | 99,5    | 100,0            |                       |
| Missing 8 DK   | 4         | ,3      |                  |                       |
| 9 NA           | 4         | ,3      |                  |                       |
| Total          | 8         | ,5      |                  |                       |
| Total          | 1500      | 100,0   |                  |                       |

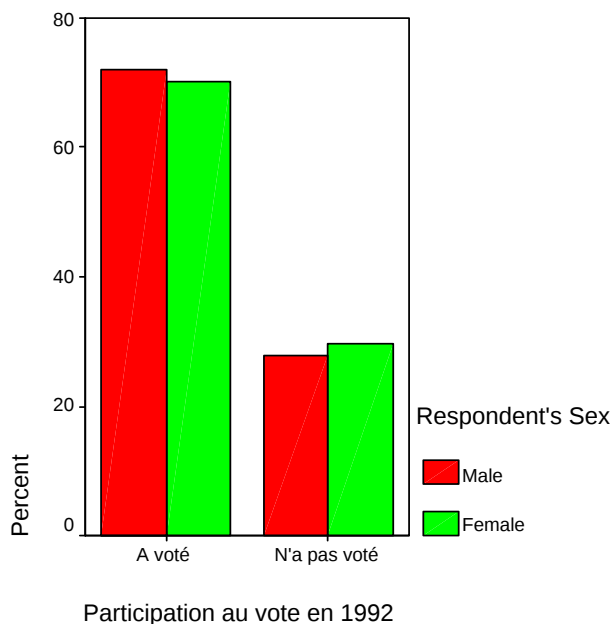
Pour la suite de l'analyse, nous allons recoder la variable pour mettre dans une unique catégorie toutes les données manquantes, sans distinction. Le tableau obtenu est le suivant (tableau 4a):

**Tableau 4a. Participation au vote en 1992**

|         |                      | Frequency | Percent | Valid Percent |
|---------|----------------------|-----------|---------|---------------|
| Valid   | 1 A voté             | 1032      | 68,8    | 71,1          |
|         | 2 N'a pas voté       | 420       | 28,0    | 28,9          |
|         | Total                | 1452      | 96,8    | 100,0         |
| Missing | 9 Données manquantes | 48        | 3,2     |               |
| Total   |                      | 1500      | 100,0   |               |

(Notons que les données ne nous disent pas si les répondants ont voté ou pas, mais plutôt s'ils ont déclaré avoir voté. Nous ferons ce rappel de temps en temps....)

**Effet du sexe.** Le graphique 1 montre que les hommes et les femmes de cet échantillon se comportent à peu près de la même façon. En effet, 72,1 des hommes ont déclaré avoir pris part au vote, contre 70.3 % des femmes, une différence minime.

**Graphique 1. Participation des hommes et des femmes au vote en 1992.**

**Effet du niveau d'éducation.** L'effet du niveau d'éducation sur la participation au vote est marquant. Le tableau 5 montre en effet que parmi ceux qui ne détiennent pas de diplôme universitaire, 65,5 % déclarent n'avoir pas participé au vote, alors que près de 90 % de ceux et celles qui ont un diplôme universitaire déclarent avoir voté.

**Tableau 5. Participation au vote en 1992 en fonction de la détention ou non d'un diplôme universitaire**

|                |                     |                         | Participation au vote en 1992 |                |        |
|----------------|---------------------|-------------------------|-------------------------------|----------------|--------|
|                |                     |                         | 1 A voté                      | 2 N'a pas voté | Total  |
| College Degree | 0 No College degree | Count                   | 730                           | 385            | 1115   |
|                |                     | % within College Degree | 65,5%                         | 34,5%          | 100,0% |
|                | 1 College degree    | Count                   | 301                           | 34             | 335    |
|                |                     | % within College Degree | 89,9%                         | 10,1%          | 100,0% |
| Total          |                     | Count                   | 1031                          | 419            | 1450   |
|                |                     | % within College Degree | 71,1%                         | 28,9%          | 100,0% |

**Effet de la religion.** On constate que les divers groupes religieux ont tendance à déclarer qu'ils se sont prévalu de leur droit de vote à des degrés divers, mais que les différences ne sont pas majeures (pas aussi grandes que l'effet de l'éducation par exemple). Le tableau 6 donne les pourcentages pour les divers groupes, qui varient entre 64,4 % pour ceux et celles qui se déclarent sans religion, à 72,4 % pour les catholiques ainsi que pour les groupes religieux autres.

**Tableau 6. Religious Preference \* Participation au vote en 1992 Crosstabulation**

|                      |              |                               | Participation au vote en 1992 |                | Total  |
|----------------------|--------------|-------------------------------|-------------------------------|----------------|--------|
|                      |              |                               | 1 A voté                      | 2 N'a pas voté |        |
| Religious Preference | 1 Protestant | Count                         | 668                           | 266            | 934    |
|                      |              | % within Religious Preference | 71,5%                         | 28,5%          | 100,0% |
|                      | 2 Catholic   | Count                         | 233                           | 89             | 322    |
|                      |              | % within Religious Preference | 72,4%                         | 27,6%          | 100,0% |
|                      | 3 Jewish     | Count                         | 20                            | 9              | 29     |
|                      |              | % within Religious Preference | 69,0%                         | 31,0%          | 100,0% |
|                      | 4 None       | Count                         | 87                            | 48             | 135    |
|                      |              | % within Religious Preference | 64,4%                         | 35,6%          | 100,0% |
|                      | 5 Other      | Count                         | 21                            | 8              | 29     |
|                      |              | % within Religious Preference | 72,4%                         | 27,6%          | 100,0% |
| Total                |              | Count                         | 1029                          | 420            | 1449   |

|                               |       |       |        |
|-------------------------------|-------|-------|--------|
| % within Religious Preference | 71,0% | 29,0% | 100,0% |
|-------------------------------|-------|-------|--------|

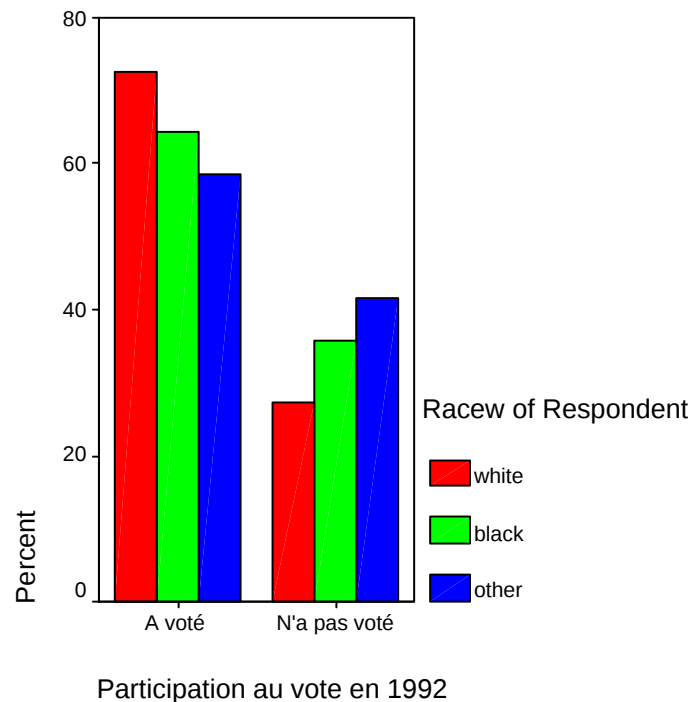
**L'effet de la classification raciale.** On observe ici une différence majeure entre les groupes définis par la classification américaine en termes raciaux. Si les blancs disent avoir voté à 72 %, les noirs à 64 %, et les membres des autres groupes à 58 %, tel qu'illustré par le tableau 7.

**Tableau 7. Participation au vote en 1992 en fonction de la classification raciale**

|                     |                              | Participation au vote en 1992 |                | Total  |        |
|---------------------|------------------------------|-------------------------------|----------------|--------|--------|
|                     |                              | 1 A voté                      | 2 N'a pas voté |        |        |
| Racew of Respondent | 1 white                      | Count                         | 893            | 337    | 1230   |
|                     |                              | % within Racew of Respondent  | 72,6%          | 27,4%  | 100,0% |
|                     | 2 black                      | Count                         | 101            | 56     | 157    |
|                     |                              | % within Racew of Respondent  | 64,3%          | 35,7%  | 100,0% |
|                     | 3 other                      | Count                         | 38             | 27     | 65     |
|                     |                              | % within Racew of Respondent  | 58,5%          | 41,5%  | 100,0% |
| Total               | Count                        | 1032                          | 420            | 1452   |        |
|                     | % within Racew of Respondent | 71,1%                         | 28,9%          | 100,0% |        |

Ceci est illustré par le graphique 2.





- V. Choisissez une association statistique observée à la question IV, et discutez en détail dans quelle mesure elle est vraie pour l'ensemble de la population dont provient cet échantillon, à supposer que ce soit un échantillon aléatoire.

Choisissons l'association entre le niveau d'éducation et le fait de voter ou pas. Nous avons vu plus haut que 65 % de ceux qui n'avaient pas de diplôme universitaire avaient voté, alors que 90 % des détenteurs de diplôme s'étaient prévalus de ce droit (en supposant que les déclarations d'avoir voté sont conformes au comportement actuel...ce qui n'est peut-être pas le cas). Pour savoir si cette différence est généralisable (elle semble bien l'être vu la taille de l'écart !) il faut calculer le Chi deux. L'hypothèse nulle est qu'il n'y a pas de différence, et l'hypothèse alternative est qu'il y en a.

Le Chi deux a une valeur de 74, qui donne un niveau de signification plus petit que 0,000. Ceci signifie qu'on peut accepter l'hypothèse alternative (à l'effet qu'il y a une différence entre les deux groupes) avec une probabilité presque nulle de se tromper.

**Note :** si vous faites le test du Chi deux pour la variable Sexe, vous obtiendrez un niveau de signification de 0,83, qui signifie que si vous retenez l'hypothèse alternative, vous aurez 83 % de chances de vous tromper !! Vous retenez donc l'hypothèse nulle (le sexe n'a pas d'effet sur le fait de voter ou pas).

Si vous aviez retenu les variable participation au vote et race, le Chi deux serait de 9,898, avec un niveau de signification de ,007 (Bond. James Bond). Vous retenez donc l'hypothèse alternative (il y a un lien au niveau de toute la population) puisque cette probabilité est plus petite que 5 %.

- VI. Choisissez un échantillon aléatoire de 100 personnes, et calculez l'âge moyen des individus de cet échantillon ainsi que le pourcentage d'entre eux qui ont voté en 1992. Sur la base de

cet échantillon, écrivez deux énoncés pour estimer l'âge moyen de la population, puis le pourcentage de gens qui ont voté aux élections de 1992 au niveau de toute la population.

Un échantillon de près de 100 personnes a été choisi. Le nombre exact choisi s'est avéré être 99. Leur âge moyen est de 48,62 ans, et 64,9 % d'entre eux ont voté. Sur cette base, nous pouvons faire les énoncés suivants :

**Estimé de l'âge de la population.**

Sur la base des données provenant d'un échantillon aléatoire de 99 personnes, nous estimons, avec un niveau de confiance de 95 %, que l'âge moyen de la population dont provient cet échantillon se situe quelque part entre 45,13 et 52,10 ans.

Ou encore

En partant d'un échantillon aléatoire de 99 personnes, nous estimons que l'âge moyen de la population est de 48,62 ans, avec une marge d'erreur de  $\pm 3,48$  ans. La probabilité d'erreur est de 5 %.

**Estimé du pourcentage de ceux qui ont voté**

Sur la base d'un échantillon aléatoire de 99 personnes, nous estimons que le pourcentage de personnes se prévalant de leur droit de vote se situe autour de 65 %, avec une marge d'erreur de  $\pm 9$  %, 19 fois sur 20.

(Au lieu de 19 fois sur 20, on peut aussi dire :

avec une probabilité d'erreur de 5 %

ou

avec un niveau de confiance de 95%.

## RÉFLEXIONS CRITIQUES SUR L'USAGE SOCIAL DES MÉTHODES QUANTITATIVES

### I. Le palmarès des écoles secondaires de l'Actualité

La discussion en classe va porter sur ces questions. Veuillez lire attentivement les textes proposés (le texte de l'Actualité ainsi que les textes critiques suggérés) et réfléchir aux questions suivantes.

1. Quel est le concept principal qui est au centre de la recherche dont fait état l'Actualité ? Quels sont les autres concepts (secondaires) qui sont aussi mesurés ?
2. Quelles sont les variables qui sont données dans le palmarès ? Sont-elles indiquées dans le texte ?
3. Quels sont les indicateurs utilisés pour mesurer ces concepts ? (pour chaque concept faites une liste des indicateurs utilisés)
4. Quels sont les arguments de nature méthodologique qui remettent en question ce palmarès comme outil de connaissance de la réalité scolaire au Québec ? Résumez les principales critiques faites au palmarès.
5. Quelles réponses donneriez-vous à ces critiques, après avoir relu le texte de l'Actualité ?
6. Compte tenu de ces critiques et des réponses qui leur sont apportées, quelle est, selon vous, la valeur de ce palmarès comme outil de connaissance ? (En d'autres termes : quelles sont les conclusions de l'étude que l'on peut prendre telles quelles, et quelles sont celles qu'il faut remettre en question ?)

### II. Le concept de Seuil de la Pauvreté

Lire le texte de Ian Hacking « Façonner les gens : Le seuil de pauvreté » tiré de : *L'ère du Chiffre : systèmes statistiques et traditions nationales*, sous la direction de J-P Beaud et J-G Prévost, Sainte-Foy, Presses de l'Université du Québec, 2000.

La discussion en classe portera aussi sur ce texte.

# Statistiques

## Statistiques descriptives

Ensemble de méthodes et de techniques qui visent à résumer des données numériques en quelques nombres, tout en saisissant les caractéristiques les plus importantes et les plus pertinentes. Une partie de l'information est perdue dans le processus.

### Mesures de tendance centrale

Elles répondent à la question: Quelles sont les valeurs les plus représentatives de l'ensemble des données ?  
Moyenne, Médiane, Mode.

### Mesures de dispersion

Elles répondent à la question : Quelle est la dispersion, ou l'éparpillement des données ? Sont-elles concentrées autour de leur tendance centrale, ou bien dispersées sur une grande étendue ?  
Écart type, variance, étendue.

### Mesures de position

Elles répondent à la question: Comment se positionnent les données individuelles par rapports aux autres ?  
Percentiles, deciles, quartiles.

### Fréquences et pourcentages

Mesures qui répondent à la question: Comment les données sont-elles distribuées sur les différentes catégories d'une variable qualitative, ou sur les valeurs d'une variable discrète ?

### Mesures d'association

Elles répondent à la question : Si on connaît le score d'un individu sur une variable, dans quelle mesure peut-on prédire son score sur une autre variable ?  
Coefficient de corrélation ( $r$ ), Khi deux.

## Inférence statistique

Ensemble de méthodes et de techniques qui visent à inférer des caractéristiques numériques d'une population lorsqu'on n'en connaît qu'un échantillon. L'inférence implique toujours une marge d'erreur ainsi qu'une probabilité d'erreur. Quand elle est fondée sur un échantillon représentatif, l'inférence a de meilleures chances de donner des résultats proches de la réalité.

### L'estimation

Elle consiste à proposer la valeur d'un **paramètre** (mesure prise sur une population) quand seule la **statistique** (mesure prise sur un échantillon) est connue. Les sondages d'opinion sont toujours fondés sur des estimations : Une enquête est menée sur un échantillon, et les résultats généralisés à la population toute entière, avec une marge d'erreur et une probabilité d'erreur.

### Les tests d'hypothèses

Ils ont pour objectif de déterminer s'il faut accepter comme vraisemblables des suppositions que l'on fait sur une population, ou de les rejeter parce qu'elles sont invraisemblables. Le processus logique est l'opposé de celui de l'estimation. On fait une supposition sur la valeur d'un paramètre. Sur cette base, on prédit qu'un échantillon aléatoire devrait probablement tomber dans un certain intervalle de valeurs. On mesure ensuite la statistique sur l'échantillon. Si elle tombe dans l'intervalle prévu, on se dit que l'hypothèse n'est pas invraisemblable. Si elle tombe en dehors de l'intervalle, on se dit que l'hypothèse est invraisemblable, et on adopte alors une hypothèse alternative qui aura été précisée dès le début du processus.

# Statistiques inférentielles

Ensemble de méthodes et de techniques statistiques visant à inférer les caractéristiques d'une population (i.e. un paramètre) à partir de la connaissance d'un échantillon (i.e. une statistique)

## Estimation

- On part d'un échantillon. Une **statistique** est mesurée.
- On **généralise** à l'ensemble de la population (i.e. on estime le **paramètre**), en prenant en considération que :
  - a) notre estimé est approximatif (il y a donc une **marge d'erreur**)
  - et que
  - b) notre estimé pourrait être complètement faux, ce qui se produirait si notre échantillon était exceptionnellement différent de la population (il y a donc une **probabilité d'erreur**)

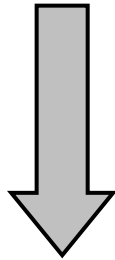
## Tests d'hypothèses

- On propose une **hypothèse** à propos de la valeur d'un **paramètre**.
- Sur la base de cette hypothèse, on **prédit** que la **statistique** correspondante va tomber dans un intervalle entourant la valeur supposée (soit dans la **zone d'acceptation**).
- Ensuite, on mesure la statistique, et on constate si elle tombe ou pas dans la zone d'acceptation prédite.
- **On tire une conclusion** :  
Si la statistique tombe **dans** l'intervalle prédit (i.e. la zone d'acceptation), on accepte l'hypothèse **comme étant probablement vraie**.  
Si elle tombe **en dehors de l'intervalle prédit** (i.e. dans la zone de rejet) on rejette l'hypothèse en se disant qu'elle est **probablement fausse**.

# Validation d'hypothèses

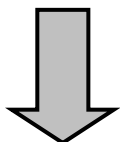
## LA LOGIQUE ... ... ET LES ÉTAPES CONCRÈTES

Une hypothèse est formulée au sujet de la valeur d'un paramètre



Sur la base de cette hypothèse, on prédit la valeur de la statistique correspondante.

Une zone de rejet et une valeur critique sont déterminées



Cette hypothèse peut être justifiée par une connaissance préalable, ou par analogie avec des situations similaires. On suppose généralement que la situation étudiée ne diffère pas notablement de celle que l'on connaît. C'est pour cela que l'on nomme l'hypothèse de départ :

**HYPOTHÈSE NULLE,  $H_0$ .**

Et aussi une

**HYPOTHÈSE ALTERNATIVE, OU HYPOTHÈSE DE RECHERCHE,  $H_1$**

qui sera retenue si  $H_0$  est rejetée. Par exemple, si l'hypothèse porte sur la moyenne  $\mu$  d'une variable, nous aurons trois possibilités :

$H_0 : \mu = 34$   
 $H_1 : \mu \neq 34$

ou  $H_0 : \mu = 34$   
 $H_1 : \mu < 34$

ou  $H_0 : \mu = 34$   
 $H_1 : \mu > 34$

Raisonnement : Si l'hypothèse est vraie, l'échantillon aléatoire choisi ne devrait pas être trop différent de la population, et sa moyenne ne devrait pas trop s'écarter de celle de la population. Mais on tolère une certaine différence car l'échantillon n'est pas une copie conforme miniaturisée de la population.

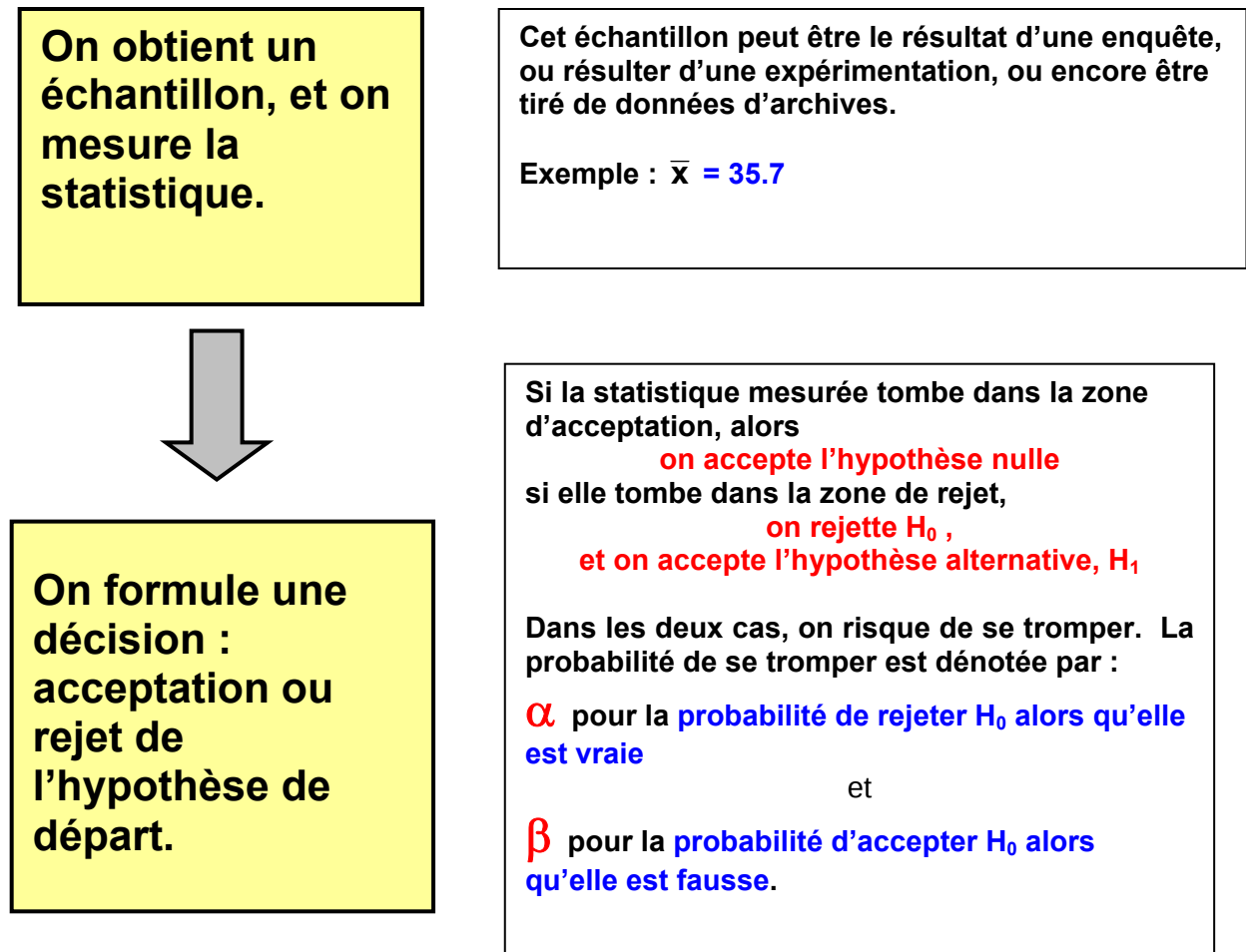
Ainsi, la prédiction fondée sur la première des hypothèses nulles ci-haut est : la **moyenne de l'échantillon,  $\bar{X}$ , devrait tomber entre 32 et 36** (i.e. on introduit une marge d'erreur de  $\pm 2$  unités max par rapport à la valeur supposée qui est 34). Cette marge d'erreur est calculée en faisant appel aux propriétés de la distribution d'échantillonnage (soit la distribution normale ou la distribution t de Student). Donc :

**Zone de rejet :**  $\bar{X} < 32$  ou  $\bar{X} > 36$

**Valeurs critiques :** 32 et 36

**Zone d'acceptation :**  $32 < \bar{X} < 36$

Si la moyenne de l'échantillon tombe dans la zone de rejet, on rejette  $H_0$  et on retient  $H_1$  comme étant fortement probable, connaissant la probabilité de nous tromper. Sinon, on se dit que l'on a pas assez de raisons de rejeter  $H_0$ .



### Remarques

1. Quand on rejette  $H_0$ , on connaît le risque que l'on prend de se tromper. En fait, c'est nous qui déterminons au départ le niveau de risque que l'on est prêt à prendre (généralement 1 % ou 5 %), et sur la base de ce niveau de risque on calcule les valeurs critiques. Donc, si on se trompe, on sait quel risque on prend exactement lorsqu'on rejette  $H_0$  , risque qu'on dénote par  $\alpha$ .
2. Mais lorsqu'on accepte  $H_0$ , on ne sait pas quelle est la valeur exacte de  $\beta$ . Tout ce que l'on sait, c'est que plus on diminue  $\alpha$ , plus on augmente  $\beta$  et vice-versa.
3. Pour ces raisons, on est sur des bases plus solides quand on accepte  $H_1$  que lorsqu'on accepte  $H_0$ . C'est pour cela que c'est  $H_1$  qui est considérée comme l'hypothèse de recherche que l'on souhaite prouver.

# Comment mesurer l'association statistique ?

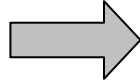
Celà dépend de l'échelle de mesure des variables

## Échelle de mesure

## Procédure pour mesurer L'ASSOCIATION STATISTIQUE

### NOMINALE VS NOMINALE

S'applique aussi aux variables quantitatives regroupées en catégories



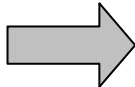
### Tableaux croisés

On compare les pourcentages horizontaux des différentes catégories de la variable indépendante. Des différences importantes indiquent une association statistique. On généralise à toute la population à l'aide du Chi deux.

**Labo 8 et Labo 13**

### NOMINALE VS QUANTITATIVE

S'applique aussi quand la première variable est quantitative regroupée en catégories



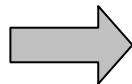
### Comparaison des moyennes

On compare la moyenne de la variable dépendante pour les diverses catégories de la variable indépendante. On généralise à toute la population à l'aide d'un test t.

**Labo 9 et Labo 12**

### QUANTIT. VS QUANTIT.

Peut quelquefois s'appliquer aux variables ordinales comportant un grand nombre de catégories



### Corrélation et régression

Le coefficient de corrélation  $r$  nous renseigne sur l'intensité de la relation et sur sa direction. La droite de régression donnée graphiquement et par une équation nous permet de prédire les scores des individus sur la variable dépendante à partir de leur score sur la variable indépendante. Ces prédictions sont toujours accompagnées d'une erreur, qui tend à être petite quand la corrélation est forte.

**Labo 10**